Analysis and Prediction of Decision Making with Social Feedback

Andrew Reed Stewart

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Mechanical and Aerospace Engineering
Advisor: Naomi Ehrich Leonard

January, 2012

Abstract

Robots that work in conjunction with humans are becoming commonplace. Some are autonomous, operating without human input, but many require supervision or direct control. In this work we suggest using *mixed teams* for decision making when robots are faced with complex tasks and human input is beneficial. Evidence that robots can be effective as peers with humans is plentiful [45, 16, 35], but new tools are needed for designing such systems.

A formal, model-based analysis of decision making in teams that share social feedback is provided. We focus on the *Two-Alternative, Forced-Choice* (TAFC) task [53] which has been widely-studied in experiments with human subjects [56, 57]. Decision makers in the TAFC task choose between two options, sequentially in time, and receive feedback on performance. This task is simple and relatively well-understood, but the predictive tools we develop and the principles that we uncover likely extend to different and more complex tasks.

Deterministic and stochastic decision-making strategies are considered. Our ability to predict behavior in teams with social feedback relies on our analysis of a stochastic soft-max choice model [28] where we reveal dependence of performance on parameters describing the task, the decision makers, and the social feedback. These tools can assist in the design of mixed teams. For example, it is possible to identify scenarios where robotic decision makers designed into a mixed team can significantly improve performance. Experiments are underway to test our hypotheses.

Relevant applications and methods of interaction are also of interest. We describe the development of a robotic testbed that supports real-time operation of multiple, robotic vehicles in a three-dimensional field. We have plans for experiments that study mixed teams working with physical robots in our testbed. By considering real-world constraints in that setting, a comprehensive, and realistic approach to studying joint decision making will continue.

Acknowledgements

First and foremost, I am thankful to Naomi Leonard, my advisor, for countless hours of her time and for providing excellent direction and encouragement throughout my years at Princeton. Princeton University, and Naomi, have provided unequaled opportunities to learn new skills, broaden my horizons, and pursue my interests. Naomi's ability to delve into new domains that range from oceanography to psychology and her success in building numerous collaborations speaks not only to her ability to perform outstanding research, but also her outgoing personality and charisma. It is simply fun to work with Naomi. Naomi's willingness to make time for students and her ability to review the details of a dense mathematical proof while simultaneously shaping our research to meet big-picture objectives is remarkable. Her attention to detail and dedication to exceptional scientific writing have made each of our manuscripts a work of art. Few people could ever direct such a diverse research program and also deliver the mathematically rigorous results that she routinely produces. I will always admire the playful curiosity and sharp intellectual focus Naomi brings to her work. It has been an honor to work with Naomi and a pleasure to be her student.

A very important collaboration with Ming Cao led to developing the theoretical framework used in this thesis. Ming is a valued colleague and friend. His ability to tackle abstract problems and build a tractable model was instrumental in setting the stage for this research. His friendship and guidance were critical during a time when I sought direction. I thank Ming for all of his support and valuable contributions that have influenced this work.

As a member of the Dynamical Control Systems Laboratory I have had the pleasure to work with a diverse group of scientists, mathematicians, and engineers. Fumin Zhang, Francois Lekien, Derek Paley and Benjamin Nabet were excellent resources as I began my graduate work. Fumin has always encouraged putting ideas into practice. His support enabled me to inherit the equipment in our labs and ultimately develop

our multi-vehicle, robotic testbed. Derek taught me the ins and outs of his thesis work, graciously sharing with me the software tools he developed and even trained me to put those tools to use. Dan Swain, Darren Pais, and Paul Reverdy have been important colleagues. Whether discussing new research directions, developing hardware for our projects in the laboratory, or just reflecting on the meaning of life as a graduate student, each has been a treasured companion.

I thank Ioannis Poulakakis, Carlos Caicedo, and Luca Scardovi for fruitful discussions and for their support of my work. I thank Stephanie Goldfarb, Kendra Cofield, Tian Shen, and George Young for providing me with helpful feedback as my research has unfolded and also for their friendship. Membership in the lab has afforded me a unique perspective and a number of rare opportunities at Princeton. Whether traveling the distance to Forrestal campus, or navigating the hallways to Von Neumann, working in our laboratories has been a special way to connect with Princeton's campus and its rich history.

Working with Andrea Nedic, Damon Tomlin, and Jonathan Cohen has been a rewarding experience and their input has directly influenced this work. I thank Andrea for sharing the models she developed and analysis of experimental data she performed. Experimental data and critical insights provided by Damon Tomlin have provided a valuable understanding and perspective. They represent the backbone of collaborations that allow us to draw conclusions about human behavior. Discussions with Jonathan Cohen and his subsequent guidance has strongly influenced directions taken. His foresight and ability to recognize promising avenues have been critical to developing hypotheses that can be tested, giving scientific merit to this work.

The Mechanical and Aerospace Engineering Department (MAE) goes beyond providing an academic and professional environment to provide high caliber education and perform cutting edge research. MAE is a community and a second home for those who frequent its halls. The warm-hearted care that Jessica Buchanan O'Leary and

Jo Ann Love provide day in and day out does more than just to keep the students on track. Candy Reed and Maureen Hickey will always go above and beyond the call of duty to assist with any task, and their kindness has done much to make MAE feel like home. I thank Jill Ray for all her time and guidance and for assisting in the preparation of my materials. I thank Marcia Kuonen, Valerie Carroll, Deborah Brown, Kathy Opitz, and Louis Rhiel for all of their assistance and for always making time for genuine discussion. Dan Hoffman, Glenn Northey, Jon Prevost, Michael Vocaturo, and Chris Zrada have each been a pleasure to work with and excellent resources for tackling tough engineering problems.

MAE faculty members Robert Stengel, Luigi Martinelli, Michael Littman, and Alexander Smits have each been very influential members of the faculty. I thank them for taking interest in my work, and for providing direction. I am thankful to Dick Miles for taking interest in my career, and for his support. I thank Phillip Holmes and Jeremy Kasdin for reading my thesis and providing constructive feedback. It has benefitted greatly from their input. Phil has been an important resource and mentor from the beginning. He has always taken the time to know my work in all of its detail. Jeremy has been a cherished professor, mentor, friend and even a rock-climbing partner over the years. He has shared with me a unique and masterful approach to solving dynamics problems, and a perspective on life that will have a continued impact on my career.

The classmates of MAE are true companions in and out of the class room. Summer softball games, weekly rock-climbing sessions, and early-morning weight-lifting routines are just a few of the extracurricular activities that have shaped my most treasured memories of Princeton. I thank Manny Stockman and Joshua Proctor for their loyal and supportive friendship. Steve Brunton, Bingni Wen Brunton, Dmitry Savransky, Tyler Groff, Lauren Padilla, Elena Krieger, Jason Kay, Marcus Hultmark,

Leo Hellstroem, and Megan Leftwich are just a few of the many important friends that have made Princeton special and will be a part of my life in years to come.

No accomplishment of mine has been made without the strong support of my parents. My mother and father are a source of inspiration and have always done everything in their power to make my dreams a reality. The education of my childhood, a unique upbringing that mixed hard labor with privileged experiences, will always be my most highly-valued degree. I thank my sister, Lindsay, for always supporting me and helping me maintain balance. I thank Emily Roche for reminding me of what matters most and for so many special memories in Princeton. If it weren't for the encouragement and support of Allan and Susan Wegner, I may never have chosen to pursue my interests with such depth. I dedicate this thesis to my dear friend, Mic, who selflessly supports me to no end. It should be noted that Mic has attended more classes at Princeton than I have. The department is also to be thanked for allowing Mic to be such an important member of the Princeton community.

This work has been funded by the Air Force Office of Scientific Research and the Office of Naval Research and this dissertation carries the number T-3239 in the records of the Department of Mechanical and Aerospace Engineering.

Contents

List of Figures

Chapter 1

Introduction

Robots are seeing increased use in a wide array of applications. Carrying out missions in the ocean, space, atmosphere, urban environments, and even working in the home, robots have proven useful in many domains and are impacting society with increasing force. Robotic systems capable of carrying out specialized tasks are developing rapidly but so must the tools with which we design the systems that go into operation. Our tools should not only be applied for the purpose of increasing the autonomy of robots, but also to assist integration so that robots are designed to make joint decisions with the humans that use them.

Developing algorithms that allow robots to complete complex tasks and adapt to changes in the environment or mission objective while recognizing patterns, weighing different options, and responding in real time is intractable for most real-world scenarios. In some cases there is no substitute for the decision-making capability of a human. In many applications there are factors that make it necessary for humans to supervise or control the operation of robots.

The need for increased autonomy of robots is not to be overlooked. It is significant and growing. Some missions have constraints that limit communication bandwidth, or create significant time delays that make it difficult for a human to be in the loop

and guide the system in real time. Examples of robots that face such constraints include planetary rovers [51, 89], subsea robots for ocean exploration including underwater gliders [6, 90, 71, 29] and other autonomous underwater vehicles [2, 3]. Even these platforms, however, are capable of sending data to a team of operators and receiving new commands at intermittent points in time during their operations. Some platforms, the Mars Rovers for example, can be re-programmed remotely.

The ability to adjust the level of a robot's autonomy is particularly useful. Unmanned systems used by the United States Air Force have this feature [36]. Take, for example, the *Predator* unmanned aircraft which is used overseas and operated remotely by pilots that fly the aircraft from a control station on the ground [86]. Platforms like the Predator can carry out some tasks autonomously (like following a specific course or maintaining an altitude), but when decisions must be made that rely on analysis of video and other sensor data, or when the quick prioritization of competing objectives is needed, human operators are critical.

When tasks are highly sensitive and a mistake can result in damage to property or loss of human life, having a human in the loop is critical. In applications such as bomb disposal, robots allow a human operator to diffuse bombs without being at the point of contact [47]. Autonomous bomb-diffusing robots that operate independently are not feasible though, particularly due to the complexity of the task. The same goes for surgical robots which require specially-trained surgeons to perform operations using a set of controls and a remote interface [25, 21]. The authors of [93] address this by introducing the notion of *task criticality* and argue that when it is high, a human should be in the loop.

In each of these examples the decision-making capability of robots limits the scope of their behavior and poses a need to include humans in the loop. It should be recognized though, that robots have specialized skills that benefit the decision-making process, even when robots lack the ability to operate autonomously. For instance,

robots can efficiently and reliably compute relevant quantities for a mission, store large amounts of information in memory, or repeatedly perform simple operations that a human would find tiresome or boring. It is also the case that robots lack some limitations that are inherent to human decision makers.

A paradigm we propose in this work is to consider some robotic decision makers to be on a peer level with their human counterparts. Building *mixed teams* of decision makers in this way, we look for ways to design integration to advantage, considering the respective strengths and weaknesses of each agent. Issues of stability, robustness of the performance to environmental effects, and social feedback all become important to consider, yet difficult to quantify. Can the system go unstable? Are there ways to monitor the decision making and adapt feedback to mitigate deterioration of performance, avoid mistakes, or take advantage of specific skills in the team? In this thesis we lay the foundation for analysis that can be used to help address these questions.

1.1 Motivation and Goals

Group decision making is at the core of the problem we consider. Robots can do many things that humans cannot, but humans have decision-making skills that are distinct and highly valuable in most applications. For example, humans are good at recognizing patterns, can adapt well to changing environments, and can prioritize competing objectives. Humans and robots each have something to offer from the standpoint of analyzing information, comparing alternatives, and maximizing an objective. Each has a distinct expertise, which motivates building a proper understanding before joining humans and robots at the supervisory level.

Proper design of systems that incorporate human input should take into consideration the fallibility of human decision-making. Stress, boredom, and fatigue are some

of the important factors [91]. There is need for a predictive capability to assist in the design of mixed decision-making teams. We propose that decision-making teams should be comprised of both human and robotic decision makers that work together. It is necessary to integrate both parties in a way that allows each to excel in the task. Doing so requires an in-depth understanding of the dynamics of decision making in mixed teams. Partitioning tasks and understanding when social feedback or the use of computer-aided (robotic) tools is beneficial are critical aspects to consider in the design of these teams.

The design of the team's network and the interconnection topology can have a dramatic effect on performance. Other driving factors include properties that define each decision maker and the *reward structure* that defines how rewards (metrics of success) are determined in the task. Proper understanding of this dependence requires new tools for design and analysis. In particular, there is need for a predictive model that describes the decision-making dynamics. A key goal of work in this area is to uncover principles that can be used to build design tools. A major hurdle is to develop an understanding of how humans make decisions in relevant, complex tasks. Human behavior is difficult to model and predict.

It is not the goal of this work to understand human decision making perfectly, nor is it to replace humans with robots. The aim, rather, is to develop a predictive model that approximates mechanisms in the human brain and can be used toward the design of mixed teams. We perform a number of model-based analyses which are used to make predictions and draw conclusions about decision-making performance. It is our goal for these predictions to aid in the design of mixed teams so that decision-making is efficient, accurate, safe and carried out in a way that takes advantage of the underlying principles of the decision-making dynamics in mixed teams.

Integrating teams of human and robotic decision makers poses a number of interesting and significant challenges. Many of them are engineering problems, but

some are philosophical and even ethical [5, 70]. Not only does this push the edge of technology, it pushes the edge of science as well.

1.2 Research Overview

We take a systematic approach to investigating human decision making through determining underlying principles and building predictive models. This is possible by focusing on the relatively well-understood *Two-Alternative, Forced-Choice* (TAFC) task, in which humans make sequential decisions among two alternatives and receive a reward after each decision [53]. We build a predictive model that describes human behavior in a complex decision-making task. The modeling effort has benefitted from collaborations with psychologists performing experiments with human subjects in the TAFC task [56, 57].

In experimental studies of the task, subjects make choices and receive rewards sequentially in time (see Chapter 3 for a detailed description of the TAFC task), and each choice is recorded along with the rewards they receive. This data is then used to analyze the behavior of human subjects, and to develop models that can approximate that behavior. By testing decision-making models against recorded data in the TAFC task, models are checked to see how well they represent the decision-making of humans. Andrea Nedic has led much of the work in model selection for the TAFC task. We use these tested models in our analyses to build a predictive capability. In particular, we focus on stochastic soft-max choice model for decision making. The soft-max choice model is a probabilistic decision-making strategy which has seen much success in fitting experimental data in the TAFC task.

Making predictions for the soft-max choice model is challenging since analytic expressions that predict how a model decision maker (in the TAFC task [53]) will converge in choice sequences have not been available prior to our analyses. One

approach is to study decision-making models using computer simulations, but this does not allow one to draw provable conclusions. In this work we derive analytic expressions that describe the steady-state behavior of the soft-max choice model in the TAFC task (as described in [53]) and provide formal proofs that are used to make predictions about performance and behavior.

By identifying a Markov Process that, under two reasonable assumptions, is equivalent to the soft-max choice model, we can predict the type of choice sequences subjects will converge to in the TAFC task. Our predictive tool is an analytic result that describes human behavior in terms of the most relevant parameters. This analytic, predictive tool allows us to show how performance depends on each parameter by deriving sensitivity of the performance explicitly.

Modeling behavior in a group, where multiple humans make decisions in the TAFC and share information, relies on understanding the dynamics of a single decision maker in the task and how social feedback (information shared among the group) influences each decision maker's behavior. We systematically build complexity into our analysis by first developing a tool for predicting the behavior of a single human subject.

We first consider two deterministic models of decision making. In our analysis of deterministic models, we consider the *Win-Stay, Lose, Switch* and a *deterministic limit of the soft-max choice model*. We show that these deterministic models are capable of reproducing similar behavior, and even replicating convergent behavior as seen in data. The desire to develop a more accurate predictive tool of a probabilistic strategy led us to consider the soft-max choice model with all of its stochasticity.

We found success with the probabilistic soft-max model and show that our predictive tool agrees well with experimental data for human subjects. Then, using an extended model proposed by Nedic et al [57], we apply our analysis to consider social feedback and develop an understanding of the behavior of a team of decision makers.

Our development provides a step towards designing automated decision makers that work as peers with human counterparts in the case that we use common models for the decision-making strategies. Using the soft-max choice model as a proxy for designed decision makers allows us to draw from experimental studies of humans working together and use the tools we develop to predict decision making in a mixed human-robot team. We can include decision makers that complement the human(s) for increasing robustness and maximizing performance. For example, we predict that in some cases a model decision maker can make optimal choice sequences in a task, but with parameter values that don't fit human behavior. In such a scenario, it may be possible to program automated, robotic decision makers to use the soft-max model as a strategy to boost performance.

It has been observed that a team of decision makers may perform better than a single human in one task, but worse in another. Predicting the dependence of performance on properties that define the task, and also the team, is critical to designing mixed teams. The predictive capability developed in this thesis can be used to inform the design process. In particular, individual decision makers can be designed to improve the team's performance. In new experiments we are testing the ability of our predictions to assist us in designing robotic decision makers integrated into a mixed team. The robotic decision makers are designed to influence the behavior of the humans so that overall performance is increased significantly. Predictions show that this is possible in a number of scenarios.

The work of this thesis has been performed with an eye on improving the operation of deployable robotic hardware. We consider robots in two categories in this thesis: first as the hardware at the point of contact, deployed remotely to carry out tasks, or explore environments, and second as decision-making agents with the ability to make computations, analyze data, store information, and draw conclusions using pre-programmed algorithms.

A significant effort has gone into developing a robotic testbed with elements that fall into both categories. Development of the testbed has involved constructing autonomous underwater vehicles that move in three dimensions within a water tank as well as supporting laboratory equipment. The result is a multi-vehicle, robotic testbed that contains robots capable of moving in three dimensions, sensing their environment, responding to human input, and also making decisions on their own. This testbed allows us to run experiments with human subjects in mixed teams that complete tasks using physical robots in relevant applications

Performing these experiments adds value to this study not just by verifying models or uncovering aspects in the decision making that cannot be observed without physical robots. Of perhaps more significant value is the ability for the implementation of studies with physical robots to drive our work toward relevant applications. In collaborations with psychologists and neuroscientists that run experiments with human subjects, decision-making tasks are designed to elicit specific behaviors, looking for fallibility in the decision making, all with an eye on uncovering processes that occur in the human brain. By focusing the design of new experiments on applications that involve physical robots, we are driving studies forward in a way that informs both the psychology and neuroscience objectives as well as those we face as engineers developing robotics technology.

1.3 Background and Related Work

The need for a multi-disciplinary approach is clear since developing tools applicable to the design of systems relying on human behavior requires a formal understanding of human behavior itself. This pushes the limits of our scientific understanding of the human brain.

An increasing effort is coming from the control theory community where some researchers are applying their tools to design integrated systems with humans and robots working together. The work of [66] and [23] looks into human supervisory control of unmanned vehicles. In [24] Dandach and Bullo are concerned with a human decision maker's speed and accuracy in sequential tasks. An interesting task that tests the ability of human subjects to make computations and estimates in a polynomial root counting game is studied in [62].

We focus our studies of human decision making around a specific task and leverage experiments carried out by our psychologist and neuroscientist collaborators who research how the human brain functions. The task we study, the TAFC task, is a sequential task in which a human subject must choose between two alternatives. The subject receives a reward following each choice. The subject continues to make choices, sequentially in time, until the experiment concludes and a payment that is proportional to the sum of individual rewards is made. Similar tasks have been studied by other behavioral scientists, psychologists, and neuroscientists and even economists. Montague and Berns [53] designed the first experiments that used a sequential decision-making task with two alternatives, and rewards administered in the same way as the task we study. The TAFC task as we consider it is equivalent to the task used by collaborators in [56] and [57].

Herrnstein, who also studied a similar sequential task with two alternatives, first described what he called the *matching law* [40]. The matching law accounts for a decision maker's tendency to make choices that converge to an equilibrium for which average rewards from competing alternatives are matched. In many settings, such a strategy is suboptimal and can be argued to be irrational. This result is strongly related to even earlier research that developed the notion of satisficing. In [73] Simon argued that humans (and other organisms) adapt their decision making to "satisfice," or make choices which are suboptimal and perhaps "good enough",

rather than converge to optimal solutions. Other related work also suggests that typical human decision makers use heuristic strategies [42, 85]. We are able to provide formal, mathematical proofs that certain decision-making models exhibit behavior that is consistent with Herrnstein's matching law.

The soft-max choice model [28] that we use in this work can be mapped to the *Drift-Diffusion Model* (DDM) [10] for the modeling of individual choices. The DDM is a stochastic model which has been widely successful in modeling decision making in a variety of contexts (in both humans and animals). The DDM has been studied widely and is quite popular for fitting data in experimental decision-making tasks. The DDM has been used in modeling perceptive tasks [64], as well as conceptual decision-making tasks [64, 63, 65, 74]. We are interested in the latter, however. First proposed by Egelman et al. [28], the DDM has seen continued use by Montague and Berns [53], and has become a well-accepted decision-making model in the context considered in this work.

Within the soft-max choice model, as it appears in [28], lies a simple reinforcement learning algorithm to model a subject's perception of reward. The structure of the model respects the constraints and role of dopamine neurons in the brain [54]. The process is therefore one that is likely to take place in the brain, considering the principals of neuronal computation. Researchers have found that feedforward control (a strategy that predicts future outcomes and makes decisions accordingly) is likely to be employed by human decision makers in some relevant tasks [14]. The learning algorithm within the soft-max model has a feedforward component and it can also be shown that, under certain parameters, it is equivalent to a feedforward inhibition model [10]. Studies of perception in human supervisory control of robotic systems (both visual and auditory) are relevant for some of the applications we consider. For examples of recent contributions see the work of Cummings [23, 26].

In designing teams of decision makers that work together, to collaborate in complex tasks, it is critical to take into consideration group effects. It is possible for groups to exhibit increased abilities like accuracy and faster response times. In [27] a perceptual task is studied in which multiple humans (a group of twenty) show a performance increase which is argued to be due to the group's increased ability to gather information about the environment. Studies of tasks that require voting show that performance depends on the rules for determining a group's decision [75]. In the social version of the TAFC task used in this work decisions are made in parallel so there are multiple outcomes (one for each subject). Some studies consider scenarios where a single decision is made according to a majority rule so that each decision maker has one vote. In [75] different majority rules are studied to see how accuracy of the decision changes.

Human-Robot Interaction (HRI) encompasses a wide range of research directions. In an effort to standardize tools for task-oriented mobile robots interacting with humans, Steinfeld indicates that important factors considered by the field include navigation, perception, management, manipulation, and social factors [76]. Social interaction for humanoid robots is a research area receiving increasing attention. Issues associated with the ability of humans and robots to relate to one another are addressed by researchers in social interaction [12, 13, 15]. Some early work in HRI focuses on teleoperation of mechanical systems that are not necessarily automated, but have robotic elements. As autonomy in systems increases, more attention must be paid to the interaction between human operator and autonomous robot. Research that pushes that boundary is closely related to the work in this thesis.

The authors of [69] point out that most methods of interaction are built under the assumption that "robotics experts" will be primary operators, but that new methods of interaction should be developed in parallel with new capabilities of robots. Methods of interaction are studied in [84, 34, 22]. In [45, 16, 35] it is argued that robots should

be treated as peers to their human counterparts. Such an approach will maximize each decision maker's ability to take advantage of the other's strengths. The notion of "sliding autonomy" is in particular quite prominent. A method for tuning the autonomy in a system where the control can be transferred among different agents is presented in [67]. The effectiveness of such systems is addressed in [44]. In much of the literature there is a focus on having humans *collaborate* with robots rather than supervise [46, 1].

Some of the first studies of decision-making dynamics of humans interacting with engineered systems in an aerospace application were performed by Stengel [88], [77]. In [77] Broussard and Stengel consider the joint system of pilot and aircraft and study stability while taking the human pilot's dynamics into consideration. The concept of principled negotiation is one that is still discussed in air traffic control research. The work of Wangerman and Stengel [88] proposed methods for implementing a distributed decision-making protocol to allow pilots to resolve conflicts in flight paths.

There are many applications of joint decision making that we find relevant. Studies of task-oriented robots that work together with human operators combine the elements of a human decision maker with capabilities of a remote robot with specialized skills. The authors of [20] studied data from a real urban search and rescue deployment at the World Trade Center. In [55] the critical aspects for the operation of rescue robotics with humans and robots working together are defined and evaluated. Of particular interest are applications that involve deployment of multiple robotic agents in a time-varying field [33, 8, 17]. Whether for gathering information or carrying out specialized tasks, applications of robotics in this area are prevalent.

1.4 Outline

The development of the tools presented in this thesis relies on a series of analyses that appear in Chapters 4 through 7. In these analyses we consider decision-making models in the TAFC task. The analyses allow us to derive an analytic, predictive too. The predictive capability is being tested in new experiments we have designed for mixed teams. These new experiments are described in Chapter 7. In Chapter 8 we discuss applications for joining mixed teams of decision makers with physical robots and outline planned experiments using our multi-vehicle robotic testbed.

Chapter 2 documents the design and development of the multi-vehicle robotic testbed. The robots that we use in our laboratory facility are submersible vehicles that can move around in three dimensions inside a large water tank. The development of the robotic testbed has been an extensive effort involving design of mechanical systems for the robotic hardware, supporting system architecture for the water tank and laboratory that houses the testbed, onboard electronics, and software as well. In Section 2.1 we motivate the need for developing physical hardware to perform experiments. Special properties of the vehicle (robot) design are contained in Section 2.3.2 and the supporting system architecture for the lab is presented in 2.3.1. A remote human interface that serves to incorporate human decision makers in the loop with the robots is also described in Section 2.3.1.

Our studies of human decision making are motivated and discussed in Chapter 3 where the TAFC task is defined. A detailed description of the TAFC task which we consider throughout this work appears in Section 3.2.1. A mathematical model of the TAFC task is given in Section 3.2.2.

Predictions of social behavior pertain to the TAFC task with an extension to include social feedback for a group of decision makers that make choices in parallel. The concept of mixing human and robotic decision makers to work together in tasks is discussed in Section 3.5 where we suggest that *mixed teams* can exhibit higher

performance when robotic decision makers are designed appropriately. The extension of the TAFC task to include social feedback is defined in Section 3.3. The models that approximate human decision-making strategies and the mechanisms that govern behavior appear in Section 3.4. The analyses of Chapters 4 through 6 refer back to the models introduced in Chapter 3.

Our preliminary analyses in Chapter 4 are applied to the deterministic decision-making models defined in Section 3.4. This served as a critical first step to developing a predictive model. In Chapter 4 we study convergence of choice sequences in the TAFC task and show that deterministic models can replicate some of the behavior as seen in data. In particular, we show that two deterministic decision-making models will converge to matching points for some of the reward structures. Matching behavior, a phenomenon discussed throughout this thesis, is defined in Section 4.1.

We successfully use the probabilistic soft-max model to continue our analysis in Chapter 5. By identifying a Markov chain which, under two assumptions given in Section 5.1, is identical to the soft-max model, we derive the steady state distribution for a model decision maker in Section 5.3. We compare results to experimental data in Section 5.5 and show that our tool is effective in predicting the behavior of a single decision maker in the TAFC task. This analysis of the soft-max choice model serves as a strong foundation for increasing the complexity of our study to include social feedback in Chapter 6.

Predicting behavior in a group of humans receiving social feedback in the TAFC is made possible by the analysis of Chapter 6. Similar to the method employed in Chapter 5, we again make use of the two assumptions in Section 5.1 and identify a Markov process for a focal decision maker. This allows us to derive an analytic expression that describes the steady-state behavior of the focal decision maker receiving feedback on the choices of other decision makers in the TAFC task. The expression, which appears in Section 6.2, serves as an important tool to describe the social influ-

ence in terms of feedback strength and the properties of each decision maker. Results from our analysis are shown to agree with experimental data through a comparison in Section 6.3.3.

Chapters 6 and 7 pertain specifically to teams of decision makers in a network, each receiving social feedback. Chapter 7 considers a mixed team with some designed (robotic) decision makers. In Chapter 7 tools developed in Chapter 6 are used to design decision makers within a network so that behavior of a focal decision maker is influenced in a significant way that has not yet been observed within groups of humans. This prediction has been used to design new experiments that are discussed in Chapter 7. The predicted gain in performance that we expect from these mixed teams is presented in Section 7.1.

In Chapter 8 we connect our studies with a real-world application of the TAFC task and discuss planned experiments that use physical robots in a study with our robotic testbed. An oil spill cleanup scenario is presented in Section 8.2.1 where the TAFC task is used to model the task of a human supervisor who collects information from robots that are mapping out an oil spill.

We aim to use decision-making teams to operate systems of robots that explore remote environments, collecting information or carrying out tasks. It is of particular interest to gain experience and understanding of how humans and physical robots work together. We therefore put the multi-vehicle robotic testbed to use in a series of planned experiments that are discussed in Chapter 8. The testbed experiments consider a team that operates our submersible robots deployed in a virtual (imposed) resource field inside the laboratory water tank. Details of the facility are in Chapter 2. A number of planned experiments appear in Section 8.2.3. In each scenario, humans are tasked with directing the robots to visit particular points of interest in the tank. This is discussed in Section 8.2.2.

The future directions which lie ahead for this research are laid out in Chapter 8 along with a summary and discussion of the contributions of this thesis.

Chapter 2

Multi-Vehicle Robotic Testbed

This chapter documents the development of a new multi-vehicle, robotic testbed which takes the current research goals of the Dynamical Control Systems Laboratory (DCSL) into consideration. First and foremost, we want to draw a connection between our work in decision making and robotics. The testbed is designed for experimental studies of collective motion, coordinated control, and decision making where humans interact with the robots to carry out tasks involving exploration of a three-dimensional environment. The testbed is comprised of mobile, submersible, robotic agents with dynamics that are simple and efficient, being well-suited for experimental studies of information collection in a three-dimensional (possibly time-varying) field.

Construction of a prototype for a new generation of robotic vehicles began in the summer of 2009. The platform we have developed, named "Beluga", is a small propeller-driven autonomous underwater vehicle with only two actuators and one sensor. It is designed to be easy to operate, and inexpensive to produce. An assembled vehicle is pictured in Figure 2.1. Four such vehicles have been constructed to date. All of the work has been performed in our facilities; from development of the embedded systems that make up our onboard computers, to the precision machining of the stream-lined fairing that houses those electronics.

DCSL, under the guidance of Naomi Leonard, has significant experience working with autonomous underwater vehicles (AUVs). Previous students have built hardware, performed experiments, and developed theory to study the dynamics of underwater gliders and propeller-driven platforms [92], [38], [9]. Contributions have not been limited, however, to the dynamics of single vehicles. DCSL pioneered research in cooperative control [31, 60] by developing control algorithms to design collective motion in multi-agent systems. This work involved development of a tank lab testbed and propeller-driven underwater vehicle named the "Grouper" [7]. In later years field experiments were run with underwater gliders deployed in Buzzards Bay, MA [94] and also Monterey Bay, CA [32, 49].

Experimental research in systems of multiple vehicles and automated coordinated control algorithms continues in DCSL. Darren Pais and Dan Swain have both performed experimental studies with MiaBot Pro ground vehicles [50] in the laboratory. Until recently, however, the lab has lacked a fully functional, robust system for experimenting with robotic vehicles in a three-dimensional field. Recent software and hardware developments have allowed us to push the limits of cost, size, and functionality of lab-scale robotic platforms. Such developments have prompted us to create a new generation of underwater vehicle platforms for experimental use in our lab.

In 2006 we took delivery of a new Tank Lab Facility at the Forrestal Campus of Princeton University. Modeled after a previous lab belonging to DCSL, the new tank lab houses a large water tank, office, workspace, and ground-vehicle facilities. This lab was designed with the intention of performing an array of experiments, some remotely operated from other parts of campus, or even collaborating institutions. Much was learned from previous experiences operating a lab with a water tank. Heating, ventilation, and air conditioning, for example, were designed specifically to support the facility. Most importantly, perhaps, is the placement of the tank in an

area with very high ceilings - thereby allowing the use of overhead cameras to monitor the tank.

Figure 2.1: Photograph of an assembled Beluga Autonomous Underwater Vehicle.

Many students have contributed to the development of Beluga and the supporting systems which make up the testbed. In the summer of 2009 undergraduate students from the Mechanical and Aerospace Engineering Department (MAE), Clayton Flanders, Richard Harris, and John Preston supported construction of a prototype vehicle. During the academic year of 2009-2010, Meghan Schoendorf of the Electrical Engineering Department helped to design and produce our first fully-functional on-board computer [68].

Continuing through the summer of 2010, MAE undergraduate students Brian Fishbein and Peter Iskaros made modifications to the prototype design and produced the machined pieces to build our first fully-functional vehicles. Valerie Karpov, an

undergraduate of the Computer Science Department, also supported development of software for video processing under the guidance of Dan Swain. During the academic year of 2010-2011 MAE undergraduates Blake Parsons and John Preston redesigned the onboard computer and brought our overhead camera system to life while also performing the first experiments in which two vehicles cooperated to find a point of maximum concentration in a virtual resource field [61].

Fellow graduate students of DCSL, Dan Swain and Paul Reverdy have also been instrumental throughout the development of this testbed. Dan Swain, through his work on a related project, RoboFish, has developed a valuable framework and library of video-processing software used to track objects in real time [82]. That software library, and Dan Swain's expertise, are a key component to the success of this system. Without Dan Swain's involvement, it would not be possible to close the loop. Paul Reverdy has been involved in nearly all aspects of the project since 2010. He has made design decisions as well as supported development of the electronics. The ability to operate experiments remotely, and to integrate decisions from a human supervisor would not be possible without Paul Reverdy's development of a human interface system.

As the work of this thesis has concluded, Paul Reverdy has assumed leadership of continued development and support of DCSL's hardware systems. He supervised the work of undergraduate students David Clifton and David Heinz during the summer of 2011. They have worked to maintain the fleet of four vehicles while also developing spare parts and incorporating new design modifications. They have also worked with the software to incorporate new control algorithms for vehicle onboard control.

The purpose of this chapter is to present key components of the system design. In Section 2.1 the need for such an experimental setup is discussed and motivated by the advantages of our approach. A number of key challenges come up when developing automated control for robots deployed underwater, and in a confined environment.

Some of those challenges are covered in Section 2.2. Many features of the individual vehicles as well as the system architecture are unique. Throughout the development process important design choices were made and are discussed with supporting motivation in Section 2.3. Specific details of the vehicle design are covered in Section 2.3.2. In an effort to develop a model for estimation and control, dynamics of the vehicle are discussed in Section 2.4. Section 2.5 lays out some future directions for the lab.

2.1 Motivation for Testbed

Theory can be validated by numerical simulation, and sometimes analytic, mathematical proofs, but conclusions drawn from such studies are limited by the models that feed them. Real-word robotics missions involve constraints, which are often overlooked or modeled in ways that improve tractability by making limiting assumptions about the system and environment. On the other hand, full-scale experiments are typically expensive, time-consuming, and make implementation of novel control paradigms risky. In most cases use of expensive hardware requires collaborations among several research institutions, in which case coming to consensus on details of an experiment requires a series of compromises and prohibits the study from employing a variety of approaches. Studies of coordinated control and human-in-the-loop algorithms are often limited to simulations and abstract settings. A physical system made up of multiple robotic agents creates a multitude of design challenges and constraints which motivate our studies and thinking about humans and robotic agents collaborating in complex tasks.

Until recently, human decision-making experiments that we have drawn from in this work have not involved robots. We are now capable of running experiments that test our predictive model of human decision-making with humans and robots

working together in the lab. The use of a laboratory facility to study decision making tasks also prompts us to consider details of real-world robotic systems that are not otherwise obvious when we limit ourselves to idealized, abstract scenarios.

Multi-vehicle systems can be studied by making approximations at various levels within the system. For example, we may consider robotic agents to be particles in the plane. Control can be designed to apply forces to each of the individual particles according to algorithms that are applied at the individual level. Conclusions can be drawn about those algorithms by directly integrating the equations of motion, yielding properties about the group-level behavior. In some cases analytical results are tractable, but typically realistic vehicle dynamics are not taken into account.

Our approach has been to construct a physical system which embodies critical real-world features and constraints, but in a way that makes implementation of automated control not only feasible, but inexpensive and accessible. This system allows us to quickly implement strategies for overcoming realistic challenges without the need to send autonomous underwater vehicles on multiple day deployments in the ocean, or unmanned air vehicles for long flights in designated fly zones. Experiments are conducted and give accurate results of performance in the presence of these real-world features and constraints which would otherwise not likely be addressed by pure simulation of a mathematical model.

Experiments take place in our laboratory facility and require less support in the form of people, energy, preparation, etc., than other options currently available. At the same time our facility provides key real-world features such as a three-dimensional environment, vehicle-level dynamics that require estimation and control, communication among agents, and collection of data in real time.

2.2 Challenges to Implementation

Our testbed consists of a group of four submersible robots which move independently in three dimensions, but are programmed with automated control algorithms. These algorithms can be designed for group-level motion that satisfies given constraints, or carries out a given task-specific objective. Several challenges have led to unique design approaches for this system. The most obvious challenge in creating a system of submersible robots is to avoid flooding onboard electronics with water, which may result in damage and even unsafe working conditions for operators. There are many ways to avoid this particular challenge, possibly the simplest being to avoid working with an underwater system altogether. The desire to deploy multiple agents which move in three dimensions for extended periods of time, however, makes the use of neutrally buoyant bodies particularly attractive. An alternative approach would be to use air vehicles, quad-rotor helicopters being a popular option, but this requires the use of lightweight bodies and also batteries - which limits the duration of use to the lifespan of a small battery.

The use of non-conducting fluid is another option. De-ionized water, or even oil, would allow the electronics to be submerged without any protective housing. Both options, however, are expensive and have significant disadvantages. De-ionized water becomes ionized after some time, and oil has a higher viscosity, is difficult to work with, and poses challenges to keep clean.

A second major challenge concerns estimating the state of robotic agents, which is typically done with a combination of sensors. The most popular sensors in use are magnetic compasses and global position systems (GPS). Neither of these are an option in this application, however, since we use a steel water tank housed in an indoor laboratory facility. The steel walls of the tank change the magnetic field and make the use of a compass ineffective. The walls of the lab also block GPS signals from being measured anywhere in the facility, let alone at the bottom of the

water tank. Further, GPS would not provide sufficient resolution to determine each vehicle's location and velocity within such a small region. Since GPS signals do not penetrate water well, in applications that require knowing vehicle positions below the surface, acoustic systems can be used in a similar fashion. However, this works best in unconfined environments and would not likely be suitable within the confines of a water tank of this size. This drives us to use a centralized estimation system that relies on overhead cameras and onboard pressure sensors to estimate the state of vehicles.

Once the state of each agent in our system is determined, control can be computed and must be sent to each vehicle. Fully decentralized systems might not require communication in this fashion, and although we have created a system that can duplicate features of a group of decentralized vehicles, we do require that commands are communicated to each of the vehicles, regardless of the situational paradigm we study. Should the need for wireless communication with each agent arise, two possible methods are available but each poses significant challenges. To do so with radio waves requires very low wavelength signals. Typical off-the-shelf technology provides low power levels that may not reliably penetrate the entire water tank. Acoustic signals can also be used, but would suffer the same issue that acoustic localization systems would within the confinement of the tank. The signal processing challenge of differentiating multiple individual signals in the presence of interference would be nontrivial. The use of tethers in the current system allows us to bypass these challenges while still providing the critical elements for a multi-agent system deployed in a three-dimensional environment.

2.3 Design

Our facility houses a water tank which holds approximately 20,000 gallons of water. It is 20 feet in diameter and 8 feet deep. A picture of the tank facility is shown in Figure 2.2. To make the best use of a limited space for deployment, our goal has been to construct small vehicles with a high level of mobility. The vehicles should be capable of visiting any location in the tank.

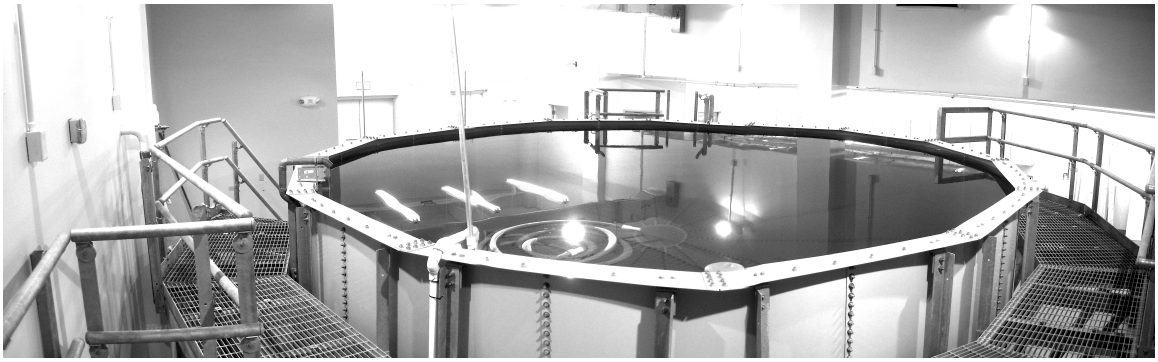


Figure 2.2: Tank facility. Steel tank approximately 20ft. diameter, 8ft. depth, filled with approximately 20,000 gallons of fresh water.

For the sake of simplicity in design, and in the interest of constructing a platform that is small yet functional, we find it convenient to limit the degrees of freedom of the vehicles while still allowing them to reach any point in the tank. The method for doing so is to build vehicles that are passively stable in pitch and roll, thereby requiring control of only four degrees of freedom. This is done by neutralizing the buoyancy of each body with a ballast below the center of buoyancy.

2.3.1 System Architecture

Computing control inputs for the vehicles requires estimating the position, orientation, and velocity of each vehicle. Depth and the vertical component of the velocity are determined by taking measurements from an on-board pressure sensor. Real-time horizontal orientation and position of vehicles are determined using a system of four

overhead cameras and video processing software. Figure 2.3 shows a single vehicle inside our lab's water tank with the flow of information depicted schematically. A computer and power supply reside alongside the tank and provide power and communication through a tether to each vehicle. The tank-side computer serves as an estimator and high-level controller, as well as a communication center for the network of vehicles.

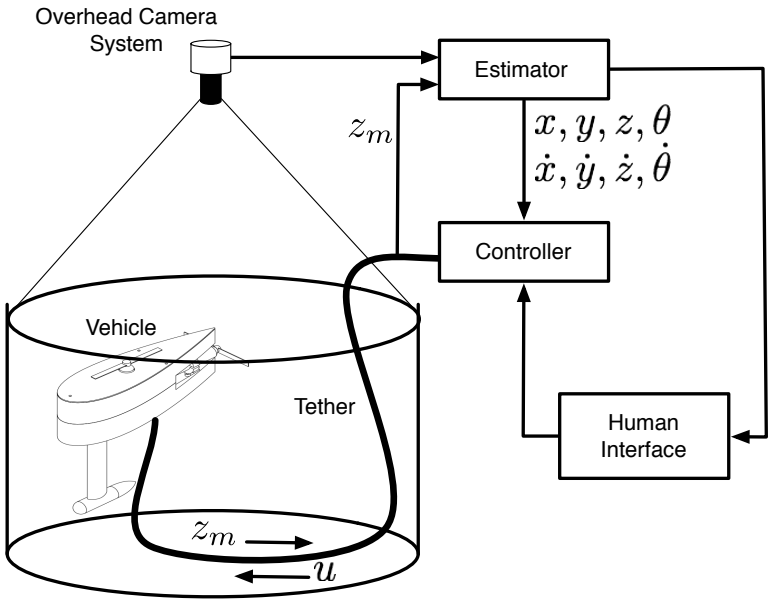


Figure 2.3: Diagram depicting information flow for a vehicle submerged in the water tank. Tether sends control input u and returns depth measurement z_m. A tank side computer performs estimation and provides control and communication. The human interface is integrated with the control system over an internet connection.

All communication is implemented using a tether with six contact elements. Each vehicle's tether carries power, as well as commands, to an onboard computer which controls the actuators. The onboard computer is capable of computing vehicle-specific control directives; i.e. maintaining depth, heading, or speed. The use of an onboard processor with significant functionality opens the door to further increase the level of autonomy. If the vehicles were equipped with batteries and wireless communication capability, the onboard computer would be ready to handle the necessary computation to make that possible.

The tank is broken into four individual quadrants, each monitored by one of the four overhead cameras (shown in Figure 2.4). To avoid tangling, we have built the tethers into the tank so that one tether connects to each quadrant. We do this by hard-wiring individual tethers to separate locations around the perimeter. Figure 2.4 depicts this layout through an overhead diagram and shows tether connection points at the perimeter of the tank for each quadrant. This setup increases separation of the tethers while still allowing vehicles to be close to one another, and visit all areas of the tank.

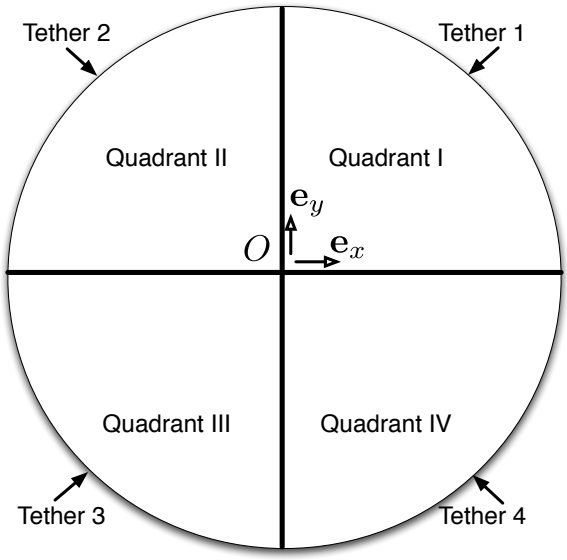


Figure 2.4: Tank layout depicting the four quadrants with tether attachment points. An inertial frame of reference is shown with origin O at the center of the tank and unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ that span the horizontal plane. Cameras are mounted from the ceiling approximately 9 ft. above the water surface over the center of each quandrant.

In some of the first studies that make use of this testbed we are running experiments that integrate a decision maker with a role equivalent to that of subjects in the two-alternative, forced-choice task. Going forward, we are taking what we have learned from our predictive capability and designing more complex paradigms for integrating the human decision making in experiments. This new experimental work is in collaboration with psychologists and behavioral scientists and is an important step

in our systematic approach to study the joint decision-making dynamics of humans and robots.

To perform experiments with humans interacting with robots in our testbed requires formal integration of a human in the loop. Figure 2.5 shows a photograph of the human interface system we have built to engage a human supervisor in decision-making tasks with the robots. In doing so, the system must provide feedback to the human and also capture the choices that the human makes.

The human is presented with four screens: three screens provide live video feeds of vehicles in the tank, and one center screen provides performance information and a control interface. All operation occurs over the internet allowing the system to be mobile; this supports the running of experiments across multiple locations. For example, human subjects can be brought into the Psychology Department to participate in experiments, while the robotic vehicles are simultaneously deployed in the tank lab. This remote operation not only allows flexibility in experimental operations, but is also representative of realistic architecture used in actual deployments when robotic platforms are deployed in remote or dangerous locations.

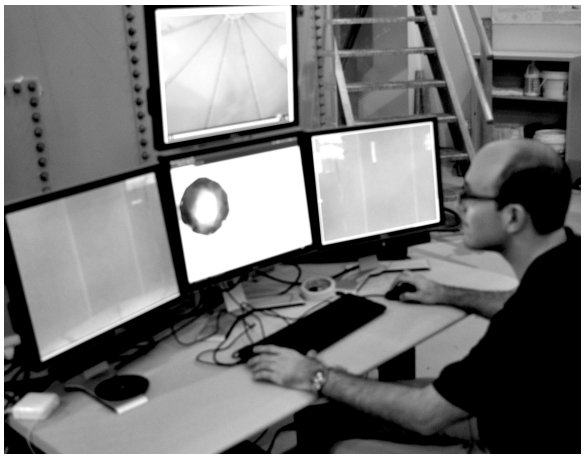


Figure 2.5: Human interface used to integrate a human decision maker with robotic vehicles. Choices are communicated to the vehicle control system, feedback is provided to the interface, and live video feeds are available for the supervisor to monitor vehicle conditions.

2.3.2 Vehicle Design

The design of Beluga was completed in three phases: (1) brainstorming and testing of individual component concepts, (2) construction of the first prototype, and (3) final modifications to the prototype and implementation of Computer Numerical Control (CNC) machining for rapid production of multiple parts. The use of CNC machining has been key to the success of building four identical vehicles in a short amount of time. Prior to building parts with the CNC, however, we tested multiple designs and iterated on each component as much as possible. Testing competing component concepts allowed us to improve the design considerably in several areas - the most significant of which applied to the method of coupling a servo to the pivot shaft of the aft, vectoring thruster. Whenever possible, we integrated components into a single module, thereby allowing ease of assembly.

Mechanical Configuration

The overall size of the vehicle is primarily driven by the onboard computer and propeller-driven actuators. While it would be possible to build a vehicle with similar functionality in a smaller package, the current design allows for the addition of battery packs, radio transmitters/receivers, and additional sensors to increase the level of autonomy and remove the tethers. The body of the vehicle is in the shape of a NACA 0012 symmetric airfoil [4]. It has an overall length of 20 in., width of 5 $\frac{3}{4}$ in., and height of 6 in. The keel hangs 8 in. below the body.

Figure 2.6 shows a drawing of the vehicle with the key components labeled. Each vehicle is equipped with two actuators. One propeller-driven thruster protrudes vertically through the body allowing vehicles to move up and down. A second vectored, propeller-driven thruster extends from the tail. This pushes the vehicle forward or backward in the horizontal plane and, when vectored, allows it to turn with a prescribed rate.

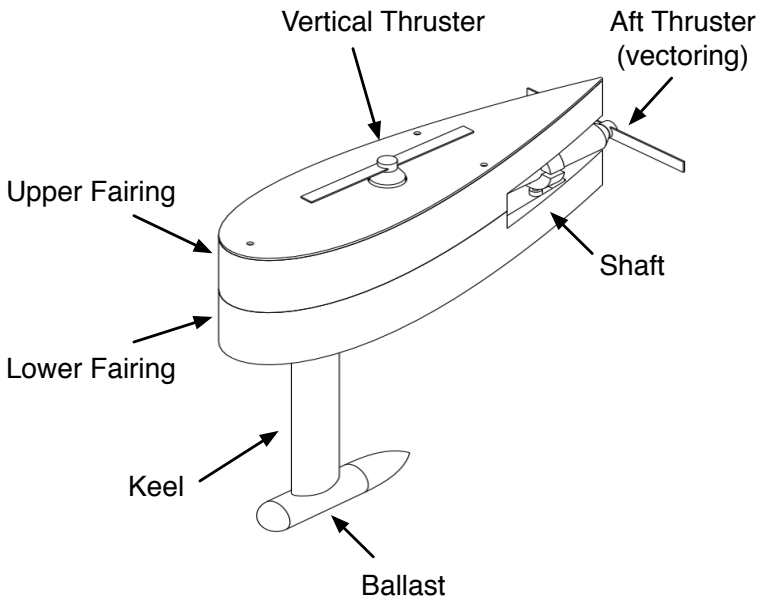


Figure 2.6: Breakout of the key vehicle components.

As shown in Figure 2.6, the body is made from an upper and lower fairing assembly. These pieces fit together in a clam shell configuration and are nearly identical, save for the fact that the upper fairing is made from low density foam, and the lower from higher density plastic, which is approximately neutrally buoyant.

In addition to the body, a ballast is made from machined brass and hangs at the end of an aluminum extrusion keel. The ballast is designed to neutralize the buoyant force of the body and is separated by a distance of approximately $L = 12$ in. from the center of buoyancy. We attempt to maximize L by making the top of the vehicle positively buoyant, and the bottom of the vehicle negatively buoyant. Figure 2.7 shows the forces due to buoyancy and gravity acting on a single vehicle, which has been rotated a small angle δ from the vertical.

This separation of the center of mass from the center of buoyancy is key to removing two of the degrees of freedom from our model. By neutralizing the buoyancy force of the body with the keel-hung ballast, we have developed vehicles that are passively stable in pitch and roll. Note that for a neutral body, the force due to buoyancy is equal in magnitude to the force due to gravity. For the forces on the vehicle shown

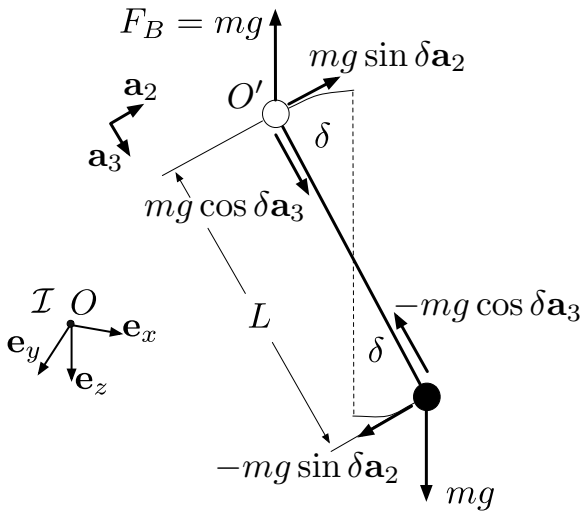


Figure 2.7: Free body diagram illustrating stability for a single vehicle. Gravitational and buoyancy forces are shown for a single vehicle which has been rolled an angle δ from the vertical. Forces are shown acting at the center of buoyancy (labeled O') and the center of mass which are separated by a distance L. The buoyancy force, F_B, acts at the center of buoyancy, the gravity force, mg (where m is the mass of the vehicle and g is the acceleration due to gravity) acts at the center of mass. The forces are given in terms of the body-fixed reference frame $\mathcal{A}$ with axes directions $\mathbf{a}_2$ and $\mathbf{a}_3$. The same diagram applies when $\mathbf{a}_2$ is replaced by $\mathbf{a}_1$ (which corresponds to a deflection in pitch).

in Figure 2.7 we can compute that given a displaced angle from the horizontal (either in pitch or roll), a resulting restoring moment is applied along the $\mathbf{a}_1$ direction ($\mathbf{a}_1 = \mathbf{a}_2 \times \mathbf{a}_3$) with magnitude

$$M = -mgL \sin \delta. \tag{2.1}$$

This tells us that the parameters driving stability with this design are L and m; namely, if we increase the mass (and also buoyancy) or the length of the keel, we should see smaller deflections δ in either pitch or roll.

Remark 1. *Note that Equation (2.1) gives us the equation of motion for a pendulum. While this analysis is accurate for determining the magnitude of the restoring moment given a deflection δ, it does not take fluid effects into account and so therefore does*

not show how disturbances to the motion in either pitch or roll are quickly damped out.

Within the vehicle body lies a chassis, which is composed of a pressure chamber to house the electronics and sensors, and a frame that houses a shaft for the aft thruster. Attention was paid toward creating a modular design. Modularity allows for ease of assembly and disassembly, as well as the possibility of modifying aspects of the design without need for altering the entire vehicle structure. The core of Beluga is made up of three main modules. They are (1) the onboard computer and front end cap, (2) the pressure chamber and main housing, and (3) the aft frame and aft end cap.

The chassis, together with the thrusters and waterproof cables, is pictured in Figure 2.8. To keep the overall length of the vehicle within a desirable range, while also allowing the vertical thruster to be located at the center of pressure of the fairing, requires a unique feature in the pressure chamber. A section of tube is welded into the chamber and the vertical thruster is secured within. This allows the onboard computer to fit forward of the thruster and for cabling to pass around to the watertight connections and pressure sensor in the rear.

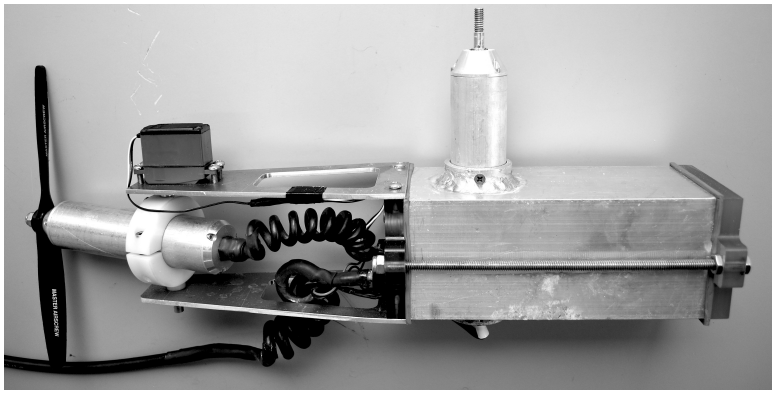


Figure 2.8: Photograph of vehicle chassis comprised of pressure chamber, end caps, aft frame assembly, and thrusters.

While typical pressure cylinders for submersible electronics are cylindrical, our approach employs the use of a square tube made from aluminum extrusion measuring

3 in. by 3 in. along the cross section. There are a number of reasons for using a square cross section chamber in our application. The first is to maximize the use of interior volume, given that we are using an onboard computer comprised of rectangular, printed circuit boards. The standard cylindrical approach requires turning cylindrical end caps on a lathe, inserting o-rings in machined grooves at a fine tolerance, and press-fitting the end caps in place. When cylindrical end caps are inserted as described, the interior volume of air is slightly compressed. This results in a force that wants to eject the end caps from the cylinder. A standard remedy is to add external braces to hold those caps in place. In our square tube, the caps butt against the ends of the chamber and so avoid compressing the interior volume.

We also find that production of square end caps and corresponding gaskets requires less time and labor given that a simple two-dimensional cut can be made using the CNC mill. Further, the square design makes it easier to encase the internal components in a machined fairing. Construction of the fairing is much simpler for our design given that we can make linear cuts through a solid body of material, rather than machining a cylindrical groove as would be necessary in the case of a traditional pressure chamber.

The aft module of the internal structure contains many of Beluga's key components in one simple, removable part. Built around the aft end cap, the aft module is made up of an aluminum frame that holds a shaft for the aft thruster. The aft thruster is mounted on this shaft to allow it to turn, thereby creating a turning moment on the entire vehicle. The angle of the aft thruster, as well as the voltage applied to the motor, controls the rate of the turn. The aft thruster angle is controlled by a submersible servo, which attaches to the aft module's frame. All water-tight connections to the pressure chamber pass through the aft end cap. This includes all cabling as well as the pressure sensor. Water-proof connectors are produced in the lab using plastic screws that have been bored out down their center. We pass each of

the contacts through the screw and embed them in epoxy. The result is a threaded unit which is mounted in the aft end cap with an o-ring to prevent leakage.

Assembly of the three main modules is achieved by a clamping mechanism. Threaded rods pass down each side of the chamber, from the forward end cap to the aft module. Two nylon insert lock nuts are tightened agains the aft end cap, evenly, to secure them in place. Immediately following assembly of the chassis, a vacuum pump fixture is used to pump air out of the chamber. That internal pressure is monitored to determine whether there are any leaks prior to introducing the vehicle to water.

Electronics

The onboard computer is a printed circuit board developed in our lab over the course of several years. It is necessary to have some computational power onboard to manage communications with a host computer to convert digital control signals into voltage output to each of the propeller-driven thrusters and to control the aft thruster's vectoring servo with a pulse-width modulated (PWM) signal. In addition to sending outputs to each of the actuators, the onboard computer converts an analog signal from the pressure sensor used for depth estimation, and sends the measurement to the host computer.

The first prototype on-board computer was developed on a breadboard using a PIC microprocessor and a small amount of peripheral hardware to support serial communications and motor control for the two thrusters. After initial functionality was achieved, we incorporated additional features, including op-amp circuitry for signal processing of the pressure sensor, and we published the design in the form of a dual printed circuit board computer. The second revision of the computer is documented in Meghan Schoendorf's senior thesis [68]. Details of the latest onboard

computer (third revision) design are described in the joint senior thesis produced by Jonathan Preston and Blake Parsons [61].

2.4 Mathematical Model of Dynamics

As mentioned previously, to keep the size small and construction simple, Beluga is equipped with only two actuators. Figure 2.9 is a free body diagram showing the chosen coordinate system with the input forces, $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ applied by the actuators. Attention was paid toward decoupling the modes of forcing; i.e. a forward thrust $\mathbf{u}_1$ applied at angle ϕ is used to drive and steer the vehicle in the horizontal plane, but should not change a vehicle's depth, z. Conversely, a vertical thrust $\mathbf{u}_2$ should not change a vehicle's heading θ or speed v. This latter constraint, however, is difficult to achieve with a single thruster providing actuation in the vertical. There is, in fact, inherent coupling between these input forces due to the limitations faced by a propeller-driven thruster.

An appropriate dynamic model will describe the dominant forces due to the surrounding fluid and also capture this coupling of the actuators applying control input. The model derived here is one such model. It is possible to include more details, or perhaps to make coarser approximations of the dynamics. Closed-loop control can make up for some uncertainty in the model, but it is important to capture the most significant driving forces. We follow a method of deriving a model for underwater vehicle dynamics presented in [37]. The fluid forces are found using a method developed by Lamb [48], which allows the fluid kinetic energy to be expressed in terms of velocity in the body frame and uses Kirchhoff's equations to relate that energy to external forces on the body. A framework for the vehicle dynamics is developed through derivation of the kinematics in Section 2.4.1 and the dynamic equations of

motion in Section 2.4.2. For more information on the notation used in deriving the dynamics see [43].

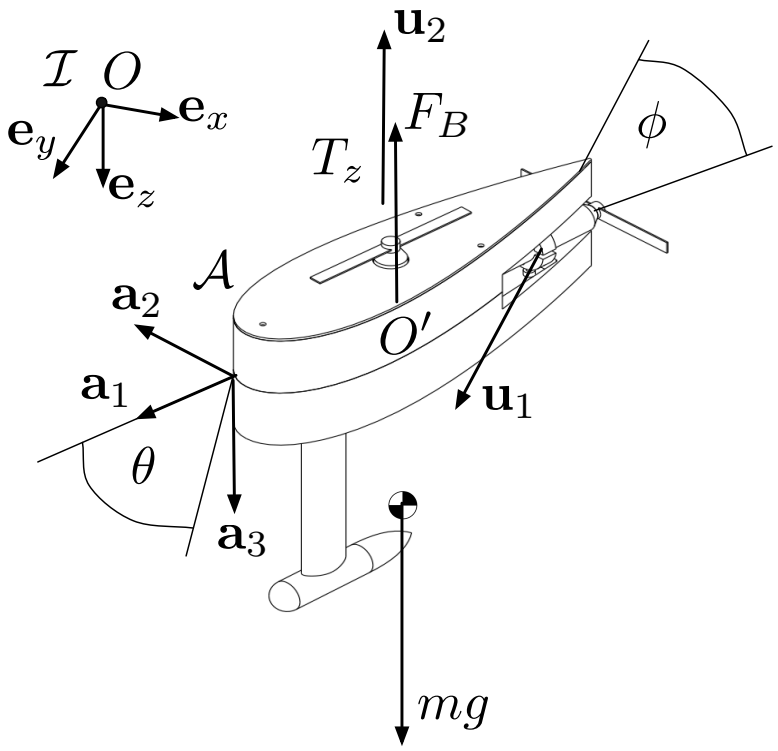


Figure 2.9: Free body diagram of underwater vehicle Beluga with coordinate system and forces applied by the actuators. $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$. The aft thruster, which vectors an angle ϕ from the vehicle centerline, is the control input $\mathbf{u}_1$. The angle ϕ lies between the axis of the aft thruster and the centerline of the vehicle spanned by $\mathbf{a}_1$. The force applied by the vertical thruster is $\mathbf{u}_2$. The force due to gravity $mg\mathbf{e}_z$ is shown acting at the center of mass. The buoyancy force $-F_B\mathbf{e}_z$ is shown acting at the center of buoyancy.

2.4.1 Kinematics

The free body diagram in Figure 2.9 shows a submerged vehicle with inertial reference frame $\mathcal{I} = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ and body frame $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$. Fluid forces and moments acting on a moving vehicle depend upon the velocity and angular rotation making it convenient to derive the kinematic equations in the body frame. We choose frame $\mathcal{A}$ to be fixed at point O' coincident with the center of buoyancy. The unit vector $\mathbf{a}_1$ spans the centerline of the vehicle and points in the forward direction, $\mathbf{a}_2$ lies in

the local horizontal orthogonal to $\mathbf{a}_1$, and $\mathbf{a}_3$ points in the local vertical as shown in Figure 2.9.

A vehicle's position in the tank is given by $\mathbf{r}_{O'/O}$. Choosing coordinates (x, y, z) to measure distances along the $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ axes we have $\mathbf{r}_{O'/O} = x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z$. The orientation of the body fixed frame $\mathcal{A}$ relative to frame $\mathcal{I}$ is given by a 3-2-1 Euler angle set $(\theta_3, \theta_2, \theta_1)^{\mathcal{I}}_{\mathcal{A}}$. Note that in Figure 2.9 we have $\theta = \theta_3$. Later in our analysis we make an assumption which allows us to describe the orientation with just one angle from the set. A corresponding transformation matrix ${}^{\mathcal{I}}C^{\mathcal{A}}$ maps vectors in frame $\mathcal{A}$ to vectors in frame $\mathcal{I}$. We choose to denote the inertial velocity of the vehicle written in terms of frame $\mathcal{A}$ by ${}^{\mathcal{I}}\mathbf{v}_{O'/O} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + v_3\mathbf{a}_3$. We also define the inertial angular velocity in body coordinates as ${}^{\mathcal{I}}\boldsymbol{\omega}^{\mathcal{A}} = \omega_1\mathbf{a}_1 + \omega_2\mathbf{a}_2 + \omega_3\mathbf{a}_3$.

The vehicles are designed to have constant distribution of mass. Although the aft thruster does pivot in order to vector the direction of thrust, the mass of that thruster relative to the overall mass is very small. It is also the case that the aft thruster pivots around a point near its own center of mass. We capture this in the following assumption.

Assumption 1. *The mass distribution of the vehicle is fixed so that the center of mass does not move within the body frame.*

$$ {}^{\mathcal{A}}\frac{d}{dt}\mathbf{r}_{P/O'} = 0 \tag{2.2}$$

where $\mathbf{r}_{P/O'}$ is the location of the vehicle's center of mass relative to point O'.

Given Assumption 1, all the information required to describe the vehicle's configuration is contained in the pair $\left(\mathbf{r}_{O'/O}, (\theta_3, \theta_2, \theta_1)^{\mathcal{I}}_{\mathcal{A}}\right)$. Using the transport equation to compute derivatives, we can express the inertial acceleration of the vehicle in the body frame via

$$
\begin{aligned}
{}^{\mathcal{I}}\mathbf{a}_{O'/O} &= \frac{{}^{\mathcal{I}}d}{dt}{}^{\mathcal{I}}\mathbf{v}_{O'/O} \\
&= \frac{{}^{\mathcal{A}}d}{dt}{}^{\mathcal{I}}\mathbf{v}_{O'/O} + {}^{\mathcal{I}}\boldsymbol{\omega}^{A} \times {}^{\mathcal{I}}\mathbf{v}_{O'/O} \qquad (2.3)
\end{aligned}
$$

where $\frac{{}^{\mathcal{A}}d}{dt}{}^{\mathcal{I}}\mathbf{v}_{O'/O} = \dot{v}_1\mathbf{a}_1 + \dot{v}_2\mathbf{a}_2 + \dot{v}_3\mathbf{a}_3$. We will use Equation 2.3 to apply Newton's Second Law in the following section.

2.4.2 Dynamics

Vehicle dynamics depend on external forces due to control inputs and also the reaction forces from the fluid that depend on the body's velocity and acceleration. In the previous section we derived the kinematic equations of motion in the body frame since here we decompose forces along the principal axes of the body frame. For a submerged body we use an *added mass matrix* to model inertial effects of the surrounding fluid. The added mass matrix allows the model to account for momentum of the fluid which accelerates as the body displaces it [37]. Viscous forces are also considered and enter into the model through external forces that we model in the subsequent section.

The kinetic energy of the fluid which has been accelerated by the vehicle can be written

$$
T_A = \frac{1}{2}({}^{\mathcal{I}}\mathbf{v}_{O'/O})^T M_A({}^{\mathcal{I}}\mathbf{v}_{O'/O}) + \frac{1}{2}({}^{\mathcal{I}}\boldsymbol{\omega}^{\mathcal{A}})^T J_A({}^{\mathcal{I}}\omega^{\mathcal{A}}) \qquad (2.4)
$$

for mass and inertia matrices M_A and J_A. The concept of fluid kinetic energy is used with Kirchhoff's equations to derive the terms in M_A and J_A. See [48] for this derivation. It is common to neglect off-diagonal terms since it is argued that most of the energy is accounted for by the diagonal terms, and closed-loop control can make up for those unmodeled dynamics.

We also choose to approximate the body as an ellipsoid. Doing so makes the mass and inertia matrix for the rigid body diagonal. This assumption is close to true for the purposes of mass distribution.

Assumption 2. *We approximate the vehicle's body to be ellipsoidal so that a point Q on its surface has position relative to O' given by $\mathbf{r}_{Q/O'} = \hat{x}\mathbf{a}_1 + \hat{y}\mathbf{a}_2 + \hat{z}\mathbf{a}_3$ which satisfies the constraint that*

$$\frac{\hat{x}^2}{a^2} + \frac{\hat{y}^2}{b^2} + \frac{\hat{z}^2}{c^2} = 1 \tag{2.5}$$

where the semi axes along the directions spanned by $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are a, b, and c, respectively.

Let M_B and J_B be the diagonal mass and inertia matrices for the rigid body. The total kinetic energy of the system is therefore

$$T = \frac{1}{2}({}^{\mathcal{I}}\mathbf{v}_{O'/O})^T(M_B + M_A)({}^{\mathcal{I}}\mathbf{v}_{O'/O}) + \frac{1}{2}({}^{\mathcal{I}}\boldsymbol{\omega}^{\mathcal{A}})^T(J_B + J_A)({}^{\mathcal{I}}\boldsymbol{\omega}^{\mathcal{A}}). \tag{2.6}$$

For the remainder of this section we will denote $M = M_B + M_A$ the total mass of the system and $J = J_B + J_A$ the total inertia.

Translational Equations of Motion

The two control forces, $\mathbf{u}_1$ and $\mathbf{u}_2$, are shown in the free body diagram of Figure 2.9. In vector form our control input is the following

$$\mathbf{u} = u_1(\cos\phi\mathbf{a}_1 + \sin\phi\mathbf{a}_2) + u_2\mathbf{a}_3 \tag{2.7}$$

where ϕ is the angle by which the aft thruster is vectored away from centerline and u_1, u_2 are magnitudes of the control forces which relate to the voltages applied to each of the motors in the thrusters. The total external force acting on the body is, however,

$$\mathbf{F} = \mathbf{u} + \mathbf{F}_E \tag{2.8}$$

where $\mathbf{F}_E$ denotes external forces other than control inputs acting on the body. In principal, $\mathbf{F}_E$ could be modeled to account for a number of external forces acting

on the vehicle. First and foremost we consider viscous fluid forces. Other forces and disturbances caused by tethers, for example, could also be included in $\mathbf{F}_E$. We assume that the buoyancy is perfectly neutral, and the force from the tether is negligible. We will limit our model of $\mathbf{F}_E$ to include the dominant fluid forces in the principal body frame directions and also make a reasonable assumption that limits the vehicle's motion to four degrees of freedom.

Assumption 3. *Let each vehicle be neutrally buoyant so that* $\mathbf{F}_B = -mg\mathbf{a}_3$. *Assume the magnitude of the restoring moment* $M_R = -mgL\sin\delta$ *is sufficient so that stability prevents any deflections in either pitch or roll thereby enforcing the constraint* $\mathbf{a}_3 = \mathbf{e}_z$.

The validity of Assumption 3 is discussed in Section 2.3 where we show that for a neutrally buoyant body with center of mass separated a distance L from the center of buoyancy (as shown in Figure 2.7) a restoring moment with magnitude $M_R = -mgL\sin\delta$ prevents the body from pitching or rolling.

Hydrodynamic forces are decomposed into the principle axes of the body frame. In the $\mathbf{a}_1$ and $\mathbf{a}_2$ directions we have drag and lift terms to consider. The external force is denoted

$$\mathbf{F}_E = -F_1\mathbf{a}_1 + F_2\mathbf{a}_2 - F_3\mathbf{a}_3 \tag{2.9}$$

where the terms F_1, F_2, and F_3 represent the component of drag and lift in the body 1,2, and 3 directions. The force magnitudes are modeled by the following:

$$F_i = (C_{D,i} + C_{L,i}\alpha_i^2)V^2, \qquad i = 1,2,3 \tag{2.10}$$

where $C_{D,i}$ is the coefficient of drag, $C_{L,i}$ the coefficient of lift, and α_i is the angle of attack, each for the i^{th} body frame direction, $i = 1,2,3$. V^2 is the magnitude of the total velocity squared, given by $({}^{\mathcal{I}}\mathbf{v}_{O'/O}) \cdot ({}^{\mathcal{I}}\mathbf{v}_{O'/O})$. The form of equation (2.10) comes from a standard approach used in airfoil theory described in [30] and [52].

The constant coefficients are determined experimentally for each vehicle in the lab. Assumption 3 allows us to approximate the force in the $\mathbf{a}_3$ direction as pure drag since the angle of attack α_3 along that direction is close to zero. We then have that $\alpha = \alpha_1 = \alpha_2 = \tan^{-1}(\frac{v_2}{v_1})$.

The external forcing, modeled in this way, is then

$$\mathbf{F}_E = -(C_{D,1} + C_{L,1}\alpha^2)V^2\mathbf{a}_1 + (C_{D,2} + C_{L,2}\alpha^2)V^2\mathbf{a}_2 + C_{D,3}V^2\mathbf{a}_3. \tag{2.11}$$

The equations of motion relating acceleration in the body frame to the forces modeled in this section are given by

$$\frac{^{\mathcal{A}}d}{dt}{}^{\mathcal{I}}\mathbf{v}_{O'/O} = M^{-1}(\mathbf{F} - {}^{\mathcal{I}}\omega^{\mathcal{A}} \times {}^{\mathcal{I}}\mathbf{v}_{O'/O}) \tag{2.12}$$

which is a system of three scalar equations. Invoking Assumptions 2 and 3, we can write out the following three equations for the variables v_1, v_2, and v_3 in the body frame:

$$\dot{v}_1 = \frac{1}{m_1}\Big(u_1\cos\phi - (C_{D,1} + C_{L,1}\alpha^2)(v_1^2 + v_2^2) + \dot{\theta}v_2\Big) \tag{2.13}$$

$$\dot{v}_2 = \frac{1}{m_2}\Big(-u_1\sin\phi + (C_{D,2} + C_{L,2}\alpha^2)(v_1^2 + v_2^2) - \dot{\theta}v_1\Big) \tag{2.14}$$

$$\dot{v}_3 = \frac{1}{m_3}\left(u_2 + (C_{D,3}\alpha^2)v_3^2\right) \tag{2.15}$$

where m_1, m_2, and m_3 are the diagonal elements of the total mass matrix $M = M_B + M_A$.

Rotational Equations of Motion

We are now left to determine the external moments acting on the body. We compute moments about the body frame origin O'. The sum of the moments, each about O'

is written

$$\mathbf{T} = \mathbf{T}_T + \mathbf{T}_z + \mathbf{T}_E \tag{2.16}$$

where $\mathbf{T}_T$ is the turning moment from the vectoring thruster, $\mathbf{T}_z$ is a torque applied by the vertical thruster, and $\mathbf{T}_E$ is an external moment acting on the body by the surrounding fluid. The vectoring thruster at angle ϕ from the vehicle centerline contributes a turning moment which can be written

$$\mathbf{T}_T = u_1 \sin\phi \mathbf{a}_3. \tag{2.17}$$

Since a propeller is a rotating airfoil, thrust is provided orthogonal to the plane of rotation, but the airfoil's lift to drag ratio, which we will call κ here, comes into play as well. The result is that the drag on the blade imposes a torque which is proportional to the thrust. So we have that the torque $\mathbf{T}_z$ applied by the vertical thruster is given by

$$\mathbf{T}_z = -\kappa u_2 \mathbf{a}_3. \tag{2.18}$$

Assumption 3 allows us to neglect moments about the $\mathbf{a}_1$ and $\mathbf{a}_2$ axes. The moment due to the surrounding fluid is modeled as

$$\mathbf{T}_E = (C_{M_0} + C_M \alpha) V^2 \mathbf{a}_3 \tag{2.19}$$

and acts purely about the $\mathbf{a}_3$ axis. Equation (2.19) is also a standard model derived in airfoil theory that uses potential flow calculations [30, 52]. Equations (2.17) - (2.19) allow us to write (2.16) as

$$\mathbf{T} = (u_1 \sin\phi - \kappa u_2 + (C_{M_0} + C_M \alpha) V^2) \mathbf{a}_3. \tag{2.20}$$

The rotational equations of motion for this system, under Assumption 3, collapse to the following scalar equation describing rotation about the $\mathbf{a}_3$ axis:

$$\ddot{\theta} = \frac{1}{J_3}\left(u_1 \sin\phi - \kappa u_2 + (C_{M_0} + C_M)(v_1^2 + v_2^2)\right), \tag{2.21}$$

where J_3 is the third diagonal element of the total inertia matrix $J = J_B + J_A$.

2.5 Future Directions

In this thesis the vehicles and testbed are used as part of a human-in-the-loop study described in Chapter 8. Development has been a collaboration among several members of the DCSL for the purpose of studying multiple-vehicle and human-in-the-loop control systems in a variety of applications. Significant effort has been made to facilitate the long-term use of this testbed in the Dynamical Control Systems Lab. Future students will be able to work with this system and integrate their code by following a simple communication protocol.

Next steps in developing and maintaining the testbed should be primarily in the area of software, modelling, and implementation. Emphasis has been put on allowing the vehicles to be controlled from any software platform. In our applications we have implemented control via Matlab and C++. In fact, the current system requires simultaneous use of Matlab and tracking software programmed in C++. Creation of a graphical user interface should be considered. Such an interface could allow new users to quickly implement control laws, change model parameters on the fly, and design virtual fields for the testbed in one organized, central piece of software. Modelling and system identification should also continue. As vehicles see continued use and as more performance data is collected, the dynamic model used for estimation and vehicle control should also improve. Higher levels of detail may also be considered

valuable; for instance, one could envision modelling the forces from tethers in the dynamics.

As with any piece of hardware, improvements in technology will prompt subsequent changes in design. One potential upgrade which should be considered is in the onboard electronic hardware. While the printed circuit board design suits the vehicles well and has proven to be stable, should additional sensors be added, or other capability desired, a change in platform is recommended. Many competing microprocessor boards are available today, some of which are less expensive and easier to operate than the chosen PIC microprocessor. One such platform is the Arduino which is open source and uses an object-oriented programming language much like C.

Should the desire to move to a fully autonomous vehicle platform arise, the current vehicles can be equipped with batteries and wireless communication devices. Off-the-shelf wireless technology may be feasible for implementation in the tank. Care should be taken, however, to choose wavelengths which are effective in penetrating the water and steel walls of the tank. Use of several antennas above the center of the tank may be a possible solution.

The multi-vehicle testbed does require continued maintenance. Some components are stronger than others and will last longer. Others, however, should be regularly renewed. Gaskets, for example, decay with time and will allow leakage. Continued use of the system will ensure that it is maintained properly. The lab should be viewed as a valuable resource for all members of the lab - not just as a tool for verification but as a system that allows students and researchers to overcome real-world constraints and stimulate interesting new research paths by grounding theory with a physical application.

Chapter 3

Decision Making Tasks and Models

3.1 Studies of Human Decision Making

Collaborations with psychologists and neuroscientists that are making headway toward understanding how the human brain functions have afforded us great success in developing a framework and foundation for our studies of decision making. By applying engineering tools, we've assisted in the development of models, performed analyses to make predictions, and have applied our results to design new experiments. This chapter presents the decision-making task and relevant models used in our approach to develop tools for the design of systems that incorporate human decision-makers in mixed teams.

We use the *Two-Alternative, Forced-Choice* (TAFC) task. In the TAFC task, a decision maker must choose one of two alternatives, sequentially in time, and receive a reward following each choice. Details of the TAFC task appear in Section 3.2.1. Our mathematical model of the TAFC task first appeared with preliminary results in [18] and again in [19] where we analyzed some of the decision-making models defined in Section 3.4.

We aim to discover underlying principles of the decision-making strategies employed by human decision makers. The TAFC task is a relatively well-understood decision-making task, which has allowed us to develop provable conclusions about human behavior. In this work we ask questions about optimality, and determine relevant parameters that influence performance. This allows us to design integration in ways that take advantage of the respective strengths and weaknesses of human and robotic decision makers.

First we consider a single human decision maker in the TAFC task and, through a model-based analysis, examine sensitivity of behavior to specific factors. Those factors can be parameters that define the task, the environment (the reward structure of the task), or properties of the individual decision makers. The task we choose is simple in that only two alternatives are considered in the action space, but the understanding we develop in this work is rich.

With the goal of determining how best to take advantage of human input, and how to design automated elements in the system, we focus on how to improve performance when it is sometimes hindered by properties of the environment and or feedback provided. A model-based approach allows us to develop a systematic understanding of the role of driving parameters and to make valuable predictions.

Since a key goal is to design mixed teams of decision makers, it is important that our approach extends successfully to draw conclusions about teams of decision makers that share information with social feedback. A number of decision-making models appear in this chapter. We analyze each of them in Chapters 4 through 6. In Chapter 5 we develop the ability to make predictions about behavior in groups with social feedback. This relies on the successful analysis of a stochastic soft-max choice model for decision making. In Chapter 5 we develop an analytic tool for predicting the behavior of a single model decision maker. The success we've had with the soft-max

model provides a foundation for extending our analysis to include social feedback in Chapter 6.

3.2 The Two-Alternative, Forced-Choice Task

In order to formally integrate human decision makers with a system of autonomous robots, it is necessary to quantify the properties of the human and develop tools for predicting behavior as well as for measuring and improving performance. The desire for analytical tools and a systematic approach has prompted us to study a well-understood task that has received much attention in psychology and behavioral science studies. We have chosen to consider human decision makers in the *Two-Alternative, Forced-Choice* (TAFC) task which was introduced by Montague and co-authors [28, 53], and has seen extensive use in human decision-making studies.

In the TAFC task, participants are required to choose one of the two alternatives presented to them. A reward is administered following each choice, and the task continues as the subject is prompted to make continued sequential choices. When the experiment concludes, a monetary award is payed to each subject with an amount that is proportional to the sum of individual rewards from each choice. There is an incentive, therefore, to maximize accumulated reward in the task.

Typically, experiments are performed in an abstract setting where the alternatives have no concrete meaning. It is possible to consider applications of the task, or to map the task to relevant real-world scenarios. In [19] we presented an application in which a single human supervisor was tasked with allocating robotic agents toward specific objectives. We have also argued that decision making in air traffic control is a relevant application to consider. If we consider air traffic control agents to be choosing whether planes land, or take off, and allow the reward for each choice to be related to the number of planes on the ground, then we have a task with two alternatives

that is history-dependent in a similar fashion to the TAFC task considered in this work. In Chapter 8 we present the design of new experiments with joint decision-making among humans and robots and detail an application where the TAFC task is a relevant paradigm for the human decision maker. We discuss a real-world example where decisions in the TAFC are given concrete meaning; the application considered in Chapter 8 integrates humans with robotic sensor platforms in an oil cleanup scenario.

Properties of the TAFC task and its origins are provided in Section 3.2.1. Specific "task types" are presented in Section 3.2.3. These specific tasks have a reward structure associated with them and each has at least one distinct property that can significantly affect a decision maker's behavior. That behavior is discussed briefly in Section 3.2.4, and in more depth in subsequent chapters. A mathematical model of the TAFC task is presented in Section 3.2.2

Each of the decision-making models that we use in our analyses throughout this thesis appears in this chapter. We consider both deterministic and probabilistic strategies for decision makers in the TAFC task. The simple, deterministic *Win-Stay, Lose Switch* model is presented in Section 3.4.1 and discussed further in Chapter 4. Another deterministic model, called the *deterministic limit of the soft-max model* is presented in Section 3.4.3 and also further analyzed in Chapter 4. The model which has seen the most success, however, in fitting experimental data is the stochastic *soft-max choice model* which we first introduce in Section 3.4.2. An analysis of this stochastic model and corresponding results for predicting performance of the behavior for a single human decision maker appear in Chapter 5.

Leveraging our understanding of the TAFC task, and extending the models used for describing a single human's behavior, we are able to draw insights about the behavior of decision makers in groups. Our motivation for considering a social task is for the purpose of designing decision-making teams – a concept introduced in Section 3.5. Our colleagues have developed an extension of the TAFC to incorporate multiple

participants. Details of the social extension, as well as an extension of the soft-max model to include social feedback, appear in [57]. The extension of the task to include social feedback is discussed in this work in Section 3.3. Our analysis develops a predictive tool using this extended model with social feedback. The model for a focal decision maker receiving social feedback about choices made by others in the task is defined in Section 3.4.4. The analysis of the extended soft-max choice model appears in Chapter 6.

3.2.1 Task Description

A decision maker in the TAFC task is required to choose between two alternatives (which we denote by A and B). Decisions are made sequentially in time, and a reward (performance measure) is received after each choice is made. The goal of a participant in the task is to maximize total accumulated reward over the duration of the task (optimize performance over the long run). In typical experiments a payment is awarded following a finite number of trials in the task. The payment is proportional to the sum of the individual rewards.

Participants do not know how the reward for a given choice is determined. The reward changes with time, and is not typically perceived to be constant for each choice. Though in this work we do not consider time-varying reward structures, the reward is a function not only of the immediate choice but also of the subject's recent history of choices [53, 39, 28], so there is implicit time dependence in the task. This dependence on past decisions is highly relevant for real-world, human-in-the-loop, decision-making problems where the state of a system under the control of a human supervisor will depend on the history of choices made.

When experiments are performed with human subjects in the TAFC task, the participant is prompted by a computer to make a choice by pressing either button A or button B on the screen. Once a choice is made and the button pushed, the

computer reports the magnitude of the reward. This describes the mechanics of one decision-making trial which occurs from each time step t up to $t+1$. The task repeats sequentially in this fashion up until completion of the experiment – at which point each participant receives payment that is proportional to the sum of rewards received over the course of the task. The duration of experiments varies among studies, and so does the length of time allotted for each of the individual choices. In [28] each experiment consisted of 250 sequential decisions. In [56, 57] each experiment consisted of 150 sequential decisions. Participants had 1.7 seconds to make a choice in each trial.

Experiments can also vary between the forced-choice and the free-choice protocols. The forced-choice protocol requires that a subject make a choice; i.e. the subject is given a fixed period of time after the prompt during which they should push one of the two buttons. If a choice is not made within that time the system uses the same choice that was made in the previous time step. In the free-response protocol the subject is allowed as much time as needed to make a choice. In [11] the authors studied both the free-choice protocol and the forced-choice protocol with a response time period of 2 seconds and .75 seconds. The authors of [56, 57] used the forced-choice protocol with a response time period of 1.7 seconds. The forced-choice protocol guarantees that a choice is made at regular intervals.

The decision maker must develop their own understanding of the reward structure and how the reward depends on their choices. They are not told (and should therefore have no knowledge) that reward depends on their recent history of choices. The subjects do, of course, have incentive to maximize their reward.

Experiments are designed to investigate various aspects of decision making. One such aspect is the effect of limited memory on performance. Memory is connected with the experimental control that defines the number of immediate past choices N used to determine the choice history that in turn determines the reward. That finite history

length is set for each experiment and can be varied. In the experiments of [11] and [57], $N = 20$ was used, whereas in the experiments of [28], $N = 40$ was used. In each case, N is sufficiently large so that participants cannot remember all of the decisions in the finite history used to determine the reward they receive. One can envision various real-world tasks which would depend on choice history to varying degrees. Some systems require many series of choices to be made before measurable changes in the environment or performance are available. Such a system would correspond to a task with very high N.

The reward structure itself, namely the way in which the reward depends upon subject choice history, is an important experimental control and has a dramatic effect on a subject's behavior. Determining how behavior depends on the properties defining the reward structure is a primary goal of experimental studies. We focus on four reward structures that are shown in Figure 3.1: *Matching Shoulders* (MS), *Rising Optimum* (RO), *Converging Gaussians* (DG), and *Diverging Gaussians* (DG). Versions of the MS and RO reward structures appear in [53, 11, 54]. The CG, DG, and RO structures are also studied by the authors of [57]. Through our model-based analysis, and in conjunction with experiments performed with human subjects, this work draws conclusions pertaining to behavior in each of the tasks of Figure 3.1.

While each reward structure shown in Figure 3.1 elicits specific behavior in the decision making, the level of difficulty is perhaps the first key factor to consider in each task. For example, the CG reward structure in Figure 3.1(c) is relatively easy for the human subject to make the optimal sequence of choices. The RO structure in Figure 3.1(b), however, is very difficult for a decision maker to find the optimal decision-making solution. This will be made more clear during an illustration of the task for these reward structures in Section 3.2.4.

Results specific to each of the reward structures appear throughout this thesis. Through a discussion of each reward structure in Section 3.2.3, and the corresponding

properties of behavior and predictions made through our analysis in Chapter 4-6, an in-depth and formal understanding of human decision making in terms of reward structure is developed for the TAFC task.

3.2.2 Task Model

In the two-alternative, forced-choice task a decision maker chooses one of two options labeled A and B. Decisions are made sequentially in time and a reward is administered following each choice. The value of each reward is determined by the reward structure, a set of curves which depend on the decision-making history. The four reward structures that we examine have been given in Figure 3.1 where r_A, the curve defining a reward for A, and r_B the curve defining a reward for choice B, are each plotted as a function of $y(t)$, the proportion of choice A in a decision maker's recent history (the last N choices).

We model the TAFC task by considering a focal decision maker, and we let the state of that decision maker be in the finite choice history defined by the N most recent decisions. Let $x(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ denote the last N choices of the decision maker ordered sequentially in time with $x_1(t) \in \{A, B\}$ denoting the decision at time t, $x_2(t) \in \{A, B\}$ the decision at time $t-1$, etc. We have that

$$x_k(t+1) = x_{k-1}(t), \qquad k = 2, \ldots, N, \quad t = 0, 1, 2, \ldots. \tag{3.1}$$

The proportion of choice A in the last N decisions at time t is

$$y(t) = \frac{1}{N} \sum_{k=1}^{N} \delta_{kA}(t) \tag{3.2}$$

where $\delta_{kA}(t) = 1$ if $x_k(t) = A$ and $\delta_{kA}(t) = 0$ if $x_k(t) = B$. Note that y can only take values from a finite set $\mathcal{Y}$ of $N + 1$ discrete values:

$$y \in \mathcal{Y} = \left\{ \frac{i}{N},\ i = 0, 1, \ldots, N \right\}.$$

The reward at time t is given by

$$r(t) = \begin{cases} r_A(y(t)) & \text{if } x_1(t) = A \\ r_B(y(t)) & \text{if } x_1(t) = B\,. \end{cases} \tag{3.3}$$

The average reward can be computed as

$$\bar{r}(y) = y r_A(y) + (1 - y) r_B(y). \tag{3.4}$$

For each value of y, this is the reward that would be received on average if the decision maker were to maintain that value of y. In our analysis we make use of the difference in the reward, finding it convenient to define

$$\Delta r(y(t)) := r_B(y(t)) - r_A(y(t)). \tag{3.5}$$

The variables $x(t)$ and $y(t)$ evolve as a decision maker makes choices A or B in the task. Through the rest of this chapter we define models of decision makers that are used in our analysis to understand how $x(t)$, and subsequently $y(t)$ will evolve (and or converge) in the TAFC task. We use both deterministic and probabilistic models in our analysis.

Each of the deterministic models we study are defined in Sections 3.4.1, 3.4.3, and 3.4.2. We use a stochastic system to model the decision-making processes in Section 3.4.2 and 3.4.4. When a probabilistic strategy governs the decision making, $x(t)$ and $y(t)$ take values according to a random process and may therefore be considered

random variables. When the decision making is deterministic, however, the variables $x(t)$ and $y(t)$ are not random.

3.2.3 TAFC Reward Structures

In each of the reward structures, the reward depends explicitly on both the current choice of A or B and the proportion, denoted by $y(t)$, of choices of A in the most recent N choices at time t. Let $i(t)$ denote, at time t, the number of A choices in the last N decisions, then $y(t) = i(t)/N$. Each reward structure is defined by two curves of reward as a function of y: in Figure 3.1 the dotted line plots $r_A(y)$, the reward received in the case that button A is pushed, and the solid line plots $r_B(y)$, the reward received in the case that button B is pushed. The average value of reward $r_A(y)y + r_B(y)(1-y)$ is the dashed curve.

The average value of reward is what the decision maker is incentivized to maximize and reward structure requires different behavior to maximize the reward received. An optimal strategy for a reward structure corresponds to quickly converge to y that maximizes the average value of reward. Without knowing the explicit structure of the reward, or knowing how much the structure depends on history (N), human decision makers may or may not successfully find the optimal solution in a task. Suboptimal behavior can happen as a result of many factors. It may be the case, for example, that a human does not explore the task enough with their choice sequences and therefore is unable to find the optimal sequence. Each task requires different levels of exploratory behavior for a decision maker to find the highest rewards allowable by the structure.

Each reward structure has specific properties that can influence a decision maker's behavior. The structures have been designed to represent interesting challenges which prompt specific behavior in a subject's decision-making space. While a typical real-word scenario is not likely to possess an idealized reward structure identical to one of those in Figure 3.1, it is likely that it would be comprised of a combination of these,

inheriting characteristics of several of the reward structures. In [87] the authors introduce the notion that these canonical curves can be considered basis functions for the space of reward curves that should be present in complex, real-world scenarios.

We consider each of the reward structures in Figure 3.1 in the abstract sense for the sake of gaining an understanding of several idealized cases. Each of these structures possess a different level of difficulty. In some of them subjects are likely to converge to a particular type of choice sequence which can be optimal or suboptimal. Interestingly, this behavior is largely predictable and we show this in Chapters 4, 5, and 6.

3.2.4 Illustration of the TAFC task

To illustrate the dynamics of the TAFC task we first consider one reward structure and walk through an example of a possible trajectory in the decision making. In this section we present an illustration for the MS, RO, and CG / DG structures, which each appear in Figure 3.1.

Matching Shoulders

Consider the *matching shoulders* (MS) reward structure of Figure 3.1(a) and suppose that the decision maker has chosen A half of the time in the last N trials of the task, i.e., $y = 0.5$. If the most recent choice made is A, then the reward is given by $r_A(0.5) = 0.35$. If the most recent choice made is B, then the reward is given by $r_B(0.5) = 0.5$. If the decision maker continues with choice sequences such that the proportion of choice A in the finite history remains at $y = 0.5$, then the average reward earned is given by the dashed curve as 0.425. A sustained choice sequence corresponding to $y = 0.5$, however, is not optimal; the maximum possible reward that one can receive corresponds to a choice sequence that achieves $y = 0.53$ (where the dashed curve in Figure 3.1(a) peaks). This can be seen by differentiating the average reward with

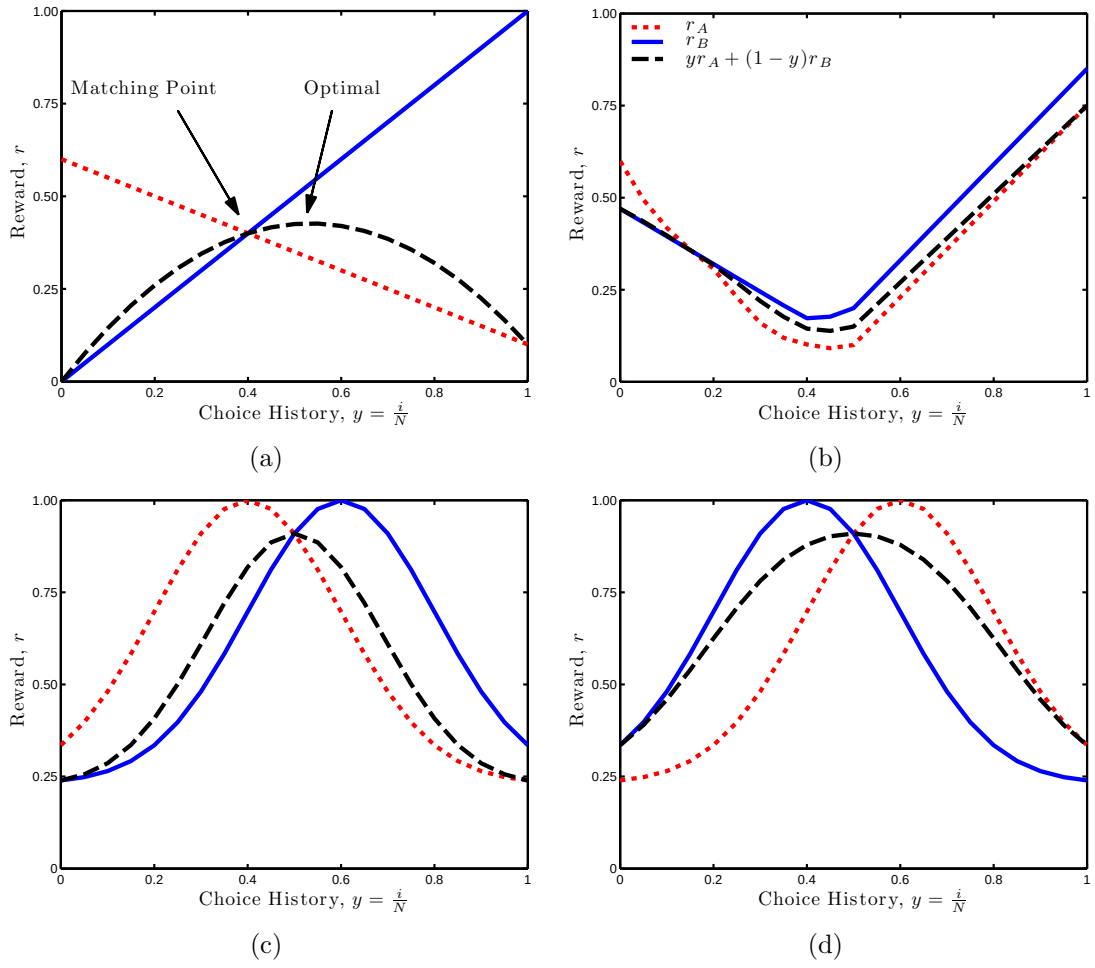


Figure 3.1: Four reward structures: (a) matching shoulders (MS), (b) rising optimum (RO), (c) converging gaussians (CG), (d) diverging gaussians (DG). In each plot the dotted line depicts r_A, the reward for choice A. The solid line depicts r_B, the reward for choice B. The dashed line is the average value of the reward. Each is plotted against y where $y = \frac{i}{N}, i = 0, 1, 2, \ldots, N$, is the proportion of choice A made in the last N decisions.

respect to y. The average reward is computed as $\bar{r}(y) = yr_A(y)+(1-y)r_B(y)$. For the MS reward structure the curves are given by $r_A(y) = k_A y + c_A$ and $r_B(y) = k_B y + c_B$. So we have $\frac{d}{dy}\bar{r}(y) = 2(k_A - k_B)y + k_B + c_A - c_B$. Solving for y that satisfies $\frac{d}{dy}\bar{r}(y) = 0$ we get $y = \frac{k_B + c_A - c_B}{2(k_B - k_A)}$. For the example of Figure 3.1(a) $k_A = -0.5$, $c_A = 0.6$, $k_B = 1$, and $c_B = 0$. In this example $\bar{r}(y)$ has a unique maximum at $y = 0.5r3$. Note that $y(t)$ is not continuous since it belongs to a finite set given by $\mathcal{Y} = \left\{\frac{i}{N}, i = 0, 1, ..., N\right\}$ and it is not always true that the value of y which maximizes $\bar{r}(y)$ is an element in that set. It is possible, however, to make choice sequences that ensure $y(t)$ is within $\frac{1}{N}$ from the optimal value.

For the MS reward structure of Figure 3.1(a), the decision maker receives a higher reward for choosing B rather than A whenever $y > 0.4$. Thus, if $y > 0.4$, the decision maker will likely make choices of B. However, continued choice of B reduces y and when $y < 0.4$, the decision maker will find that choosing A yields a higher reward than choosing B. Subsequent choices of A will increase y and the process repeats once $y > 0.4$ again. The point where the two curves intersect, here corresponding to $y = 0.4$, is called the *matching point*. Interestingly, although the matching point does not necessarily coincide with optimal choice sequences (here $y = 0.53$), it is an attractor for the models we study. This also appears to be the case for human decision makers which is in accordance with Herrnstein's matching law [41]. Evidence that human decision makers converge to choice sequences y that correspond to matching points is plentiful [28, 53, 56]. Convergence of human decision making to the matching point has been analyzed using decision-making models in [53, 18, 19, 87]. In Chapter 4 we derive analytical predictions to prove conditions under which a decision maker in the TAFC task will converge to a matching point in a reward structure.

Rising Optimum

The *rising optimum* (RO) reward structure of Figure 3.1(b) also has a matching point, but it is a more complex task since there is a local optimum at $y = 0$ and a global optimum at $y = 1$. This RO reward structure is studied in [57] with subjects who begin the task with the initial condition $y(0) = 0$. Subjects tend to spend time at the local optimum or near the matching point, but rarely find the global optimum, since to do so requires choice sequences that yield the lowest possible rewards in the task (here in the range $y = 0.4$ to $y = 0.5$). Even if a subject reaches the optimum at $y = 1$, a choice of B will yield an even higher reward than a choice of A – this will reduce y, moving the decision maker away from the optimal solution. For these reasons, the RO task is considered to be a difficult task. This difficulty level, combined with the short-term disincentive to explore, makes this a rich and interesting task for studying the effects of social feedback. For example, an important question is whether it is possible, with the right kind of feedback, that teams of humans or mixed teams of humans and robots can perform better in the RO task, e.g., find and sustain the optimal behavior, as compared to individuals who do not share information with one another. We make predictions about the implications of social feedback in the RO task in Section 7.1.

Converging and Diverging Gaussians

Figures 3.1(c) and 3.1(d) are *converging gaussians* (CG) and *diverging gaussians* (DG) reward structures, respectively. These two structures have the same form, but the curve for r_A in the CG structure is r_B in the DG structure and what is r_B in the CG structure is r_A in the DG structure. The difference between behavior in each task is significant and is inherent in the name of each. In the CG task the matching point is an attractor such that decision makers tend to converge to it, whereas in the DG

task the matching point is divergent such that decision makers tend to move away from it.

Both structures are symmetric about $y = 0.5$, which corresponds both to the matching point and to the optimal decision-making solution. The CG task is considered an easy task since the matching point, and therefore the optimal solution, is attracting. As expected, subjects on their own perform well in the CG task [57], typically finding and sustaining optimal decision sequences. However, performance was observed to degrade in the CG experiments with social feedback studied in [57]. In [80] and [79] we presented some results that agree with these experiments, those are included in Chapter 6. The social feedback (which provided information on what others were choosing or how others were performing) seemed to trigger increased exploration that led the decision maker away from the otherwise easy-to-find optimal solution. In Chapter 6 we use the model to prove the frequency of the negative impact of social feedback and show that our predicitions agree qualitatively with the experimental results of [57].

3.3 The TAFC Task in a Social Context

The authors of [56, 57] have run extensive experiments with multiple human subjects to investigate decision making in TAFC tasks with social feedback. In collaboration with this effort, we've developed a framework for studying decision making among humans in this same social context. In this section we describe the TAFC task in the social context.

In our framework we consider an arbitrary number of $M+1$ decision makers in the team. In each experiment, for the social decision-making task, five human subjects were physically isolated from one another, but they made choices for the same task at the same time. Each subject made his/her own decisions, and the corresponding

reward each subject received after each decision was calculated based only on that subject's choices according to the reward structures of Figure 3.1. After every choice, when the computer reported the reward, it also reported the current A or B choice of each of the other four subjects (choice feedback), the current reward of each of the other four subjects (reward feedback), or both the current choice and reward of the other four subjects (choice and reward feedback). In any of these social feedback conditions, each human subject could use the information reported about the four others in their own decision making. For comparison, experiments included having the five subjects perform TAFC tasks simultaneously, without feedback – this was called the *alone condition*.

The authors of [56, 57] designed their experiments with undirected feedback, that is, information passed from every individual to every other individual in the group. We study the undirected case numerically in Chapter 6. However, for the majority of our analysis, we examine the case of directed social feedback; in particular, we study the decision-making dynamics of a focal individual who receives feedback on the choices of M other decision makers, each of whom receives no social feedback.

The directed case allows us to focus on the influence of social feedback on an individual decision maker. By considering a focal decision maker, we can also investigate the influence of designed decision makers who provide feedback but have fixed strategies. We make a formal comparison between the influence of directed versus undirected social feedback in Chapter 6. In these experiments each subject's role was equal and information was shared evenly. Experiments are currently underway in which feedback is directed. We discuss those in Chapter 7. Plans are also being made to consider hierarchical decision-making paradigms.

We model the social TAFC task much like the TAFC task of Section 3.2.2. Our approach is to consider a focal decision maker in the TAFC task that receives feedback on the choices of M other decision makers. Let $x(t)$ denote the last N choices of the

focal decision maker as described in Section 3.2.2. Then $y(t)$, given by (3.2) is the proportion of choice A made by the focal decision maker in the last N trials at time t. The social feedback, dependent on the choices of the M others, appears in the decision-making model described in Section 3.4.4.

3.4 Decision-Making Models

Whenever a decision maker faces a sequential task that does not vary drastically over time, it is reasonable to believe that the behavior will converge in some way. We expect that a certain amount of learning should also occur over the task duration. Decision makers are incentivized to maximize individual rewards and must attempt to learn the value of each alternative in order to do so. Since, in the TAFC task we study, participants have no knowledge that the reward depends on their history, let alone how it does so, we have considered models from the literature [58, 28, 65, 64] that do not attempt to learn the entire structure, or the structure's specific dependence on choice history, but rather make choices in response to individual rewards received by continually learning and comparing the value of the two alternatives denoted choice A and choice B.

Each of the models in this section provides a rule for choosing A or B in the next time step, $t+1$, using information available at time t. We use both deterministic and probabilistic models in our analysis. The stochastic *soft-max choice model* is particularly successful for fitting experimental data in the TAFC task with no social feedback [28, 65, 64]. The soft-max model is defined in Section 3.4.2. To model decision making in a social context, the model is extended for the case of choice feedback [57]. The extension to the soft-max model for the TAFC task with social feedback is defined in Section 3.4.4. The *Win-Stay, Lose-Switch* (WSLS) [58] and *deterministic limit of the soft-max choice model* [19] are both deterministic decision-

making strategies. The WSLS model is defined in Section 3.4.1 and the deterministic limit of the soft-max model is defined in Section 3.4.3, respectively.

3.4.1 Win-Stay, Lose-Switch Model

The Win-Stay, Lose-Switch (WSLS) model is a simple, deterministic model of a strategy which may be employed in the TAFC task [58]. It assumes that decisions are made with information from the rewards of the previous two choices only and that a switch in choice is made when a decrease in reward is experienced.

The model chooses to repeat a choice from time t at time $t+1$ if the reward at time t is greater than or equal to that at time $t-1$; otherwise, the opposite choice is used at time $t+1$:

$$x_1(t+1) = \begin{cases} x_1(t) & \text{if } r(t) \geq r(t-1); \\ \bar{x}_1(t) & \text{otherwise,} \end{cases} \qquad t = 1, 2, 3, \ldots \tag{3.6}$$

where $\bar{\cdot}$ denotes the "not" operator; i.e., if $x_1(t) = A$ ($x_1(t) = B$), then $\bar{x}_1(t) = B$ ($\bar{x}_1(t) = A$).

Though the WSLS model can replicate common choice sequences made by human subjects, since it is deterministic, it does note accurately represent the majority of human behavior since typical choice sequences made by human participants are, at times, random. This model does, however, qualitatively represent behavior observed in experiments. In fact, it can be proved that the WSLS decision maker will converge to *matching points* in reward structures, a concept we discuss in more detail in Chapter 4 and also for stochastic models in Chapter 6.

3.4.2 Stochastic Soft-Max Choice Model

The stochastic soft-max model has been widely successful in modeling decision making in a variety of contexts (in both humans and animals) [64, 63, 65, 74]. The soft-max model, as it appears in this work to describe human decision making in TAFC task, was first proposed by Egelman et al. [28]. Having seen continued use by Montague and Berns [53], the model has since become a well-accepted decision-making model in this context.

The soft-max model prescribes the probability p_A that a subject will choose A at the next time step, using information learned about the two alternatives A and B up to time t. The probabilistic choice function is a sigmoidal function of the current state:

$$p_A(t+1) = \frac{1}{1 + e^{-\mu(w_A(t) - w_B(t))}} \tag{3.7}$$

where we have defined the probability that a subject chooses A as $p_A(t+1) := \Pr\{x_1(t+1) = A\}$. The variables w_A and w_B represent the learned value, or anticipated reward, for choices A and B, respectively. The probability p_A depends explicitly on the difference between the subject's anticipated reward w_A for pushing button A next and the subject's anticipated reward w_B for pushing button B next.

Note that w_A and w_B are independent from

The true rewards, r_A and r_B, underly the human's perception of the values w_A and w_B, although they are likely related in a complex manner. How w_A and w_B relate to r_A and r_B relies on the experience of the decision maker. This is described as a learning process later in the section.

Figure 3.2 shows p_A as a function of $w_A - w_B$ in the case that $\mu = 1$. The parameter μ is an important property of each decision maker. It determines the slope of the sigmoidal function in Figure 3.2, where one can see that larger μ implies more certainty in decision making. The parameter μ can be interpreted as the tendency

for a decision maker to explore. Smaller μ implies a decision-making rule with more random (exploratory) behavior. By inspection of (3.7) one can see that $\mu = 0$ corresponds to completely random behavior which doesn't take into account the learned reward values w_A and w_B. On the other hand, larger μ implies less randomness in the choices. As μ tends to infinity (3.7) becomes deterministic: in this case whenever $w_A > w_B$ a choice of A is made and whenever $w_A < w_B$ a choice of B is made. In [19] we analyzed this deterministic limit of the model and presented convergence results. In Section 3.4.3 the deterministic limit of this model is discussed in more detail.

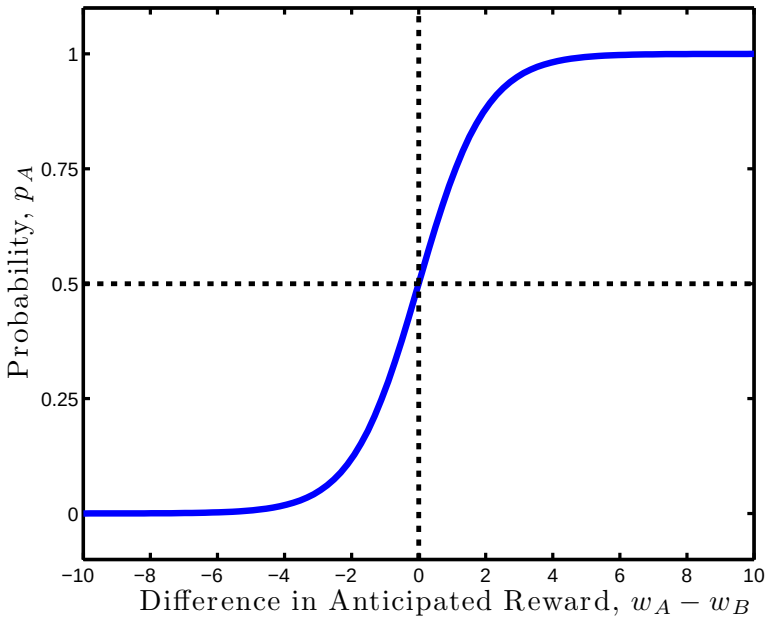


Figure 3.2: Sigmoidal function given by (3.7) representing probability of choosing A as a function of $w_A - w_B$ (plotted here with $\mu = 1$).

Since decision makers behave differently depending upon the reward structure in the task, the parameter μ can be adjusted to fit experimental data in each of the tasks. For instance, in [57] it has been found that the model best predicts a subject's choice sequences in the RO task when $\mu = 11.0$. In the CG task, however, the experimentally fitted value is $\mu = 2.50$ [57]. These values correspond to fitting parameters across all of the subjects. It is also possible to fit μ to a particular subject since each decision maker is unique. It is a goal of this work to develop a formal understanding of the

effects on performance of parameters such as μ. In the social context, a heterogeneous group of decision makers can be studied by distinguishing individuals by his or her own characteristic value of μ.

In the soft-max decision-making model proposed by Egelman et al. [28], the role of dopamine neurons was considered. In particular, the ability for dopamine neurons to code for reward prediction error [54] motivated the use of temporal difference learning theory [81] to describe the dynamic update of w_A and w_B. Let $Z \in \{A, B\}$ be the choice made at time t, then

$$w_Z(t+1) = (1-\lambda)\omega_Z(t) + \lambda r(t) \tag{3.8}$$

$$w_{\bar{Z}}(t+1) = w_{\bar{Z}}(t) \qquad t = 0, 1, 2, \ldots \tag{3.9}$$

where $\bar{\cdot}$ denotes the "not" operator. Here, $\lambda \in [0, 1]$ acts as a learning rate, controlling how the anticipated reward of choice Z at $t+1$ is affected by its value at t. Larger λ implies less "memory" since the influence of previous rewards decays. When λ is 1 there is no memory since the anticipated reward is equal to the most recent reward received.

This soft-max function (3.7) used in this decision-making model predicts individual choices in the same ways as a stochastic differential equation that describes a scalar drift-diffusion process [56, 59, 72]:

$$dz = \alpha dt + \sigma dW, \qquad z(0) = 0. \tag{3.10}$$

Here the variable z represents the accumulated evidence in favor of a candidate choice of interest, α is a drift rate representing the signal intensity of a stimulus acting on z and σdW is a Wiener process with standard deviation σ, which is the diffusion rate representing the effect of white noise.

In the context of the TAFC task with choices A and B, the drift-diffusion model can be used to predict an individual choice by allowing the drift rate α to be determined by a subject's anticipated rewards w_A and w_B. Then z, the accumulated evidence for choice A relative to choice B, evolves over time according to the stimulus and the drift. When $z(t)$ first crosses one of the predetermined thresholds $\pm\xi$ a choice is made. If $+\xi$ is crossed then choice A is made, and if $-\xi$ is crossed then choice B is made.

It can be computed using tools developed in [10] that for a drift-diffusion process defined by (3.10), the probability that z crosses one of the predetermined thresholds is given by

$$\Pr\{z = \xi\} = \frac{1}{1 + e^{-(2\alpha\xi/\sigma^2)}} \tag{3.11}$$

In [57] it is pointed out that by equating (3.7) with (3.11), we can associate $\mu(w_A - w_B)$ with $2\alpha\xi/\sigma^2$. This implies that if we associate the drift rate α in (3.10) with the difference in anticipated reward $w_A - w_B$ then we have that $\mu = 2\xi/\sigma^2$.

In our analysis we rely upon (3.7). In some cases our work refers to this decision-making model as the drift-diffusion model (for example [19, 78, 80]), though the DDM itself lacks the temporal difference algorithm (Equations (3.8) - (3.9)) needed to model the learning process which accounts for a decision maker's changing perception of the value of choices in the TAFC task.

3.4.3 A Deterministic Limit of the Soft-Max Choice Model

To gain insight into the mechanics of the stochastic choice model, we first chose to adapt the model as described in Section 3.4.2 to a deterministic one. This made the model tractable for drawing conclusions, which appear in Chapter 4.

Specifically, we study the deterministic limit of the decision rule (3.7) deduced from the soft-max choice model by letting μ in (3.7) go to infinity. Then at time $t > 0$, the subject chooses A if $\omega_A(t) > \omega_B(t)$ and B if $\omega_A(t) < \omega_B(t)$. In the event that $\omega_A = \omega_B$, we assume that humans are of an explorative nature, and thus the subject uses the opposite of the last choice. To summarize,

$$x_1(t) = \begin{cases} A & \text{if } \omega_A(t) > \omega_B(t) \\ B & \text{if } \omega_A(t) < \omega_B(t) \\ \bar{x}_1(t-1) & \text{if } \omega_A(t) = \omega_B(t) \end{cases} \quad t = 1, 2, 3, \ldots \tag{3.12}$$

where as in (3.6), $\bar{\cdot}$ denotes the "not" operator.

Now consider the dynamic update of the anticipated rewards ω_A and ω_B as modeled in (3.8)-(3.9). When a choice of Z is made the value of $\omega_{\bar{Z}}$ remains unchanged because without memory no reward information for $\bar{Z}$ is available. A more sophisticated update model, called the *eligibility trace model*, is constructed in [11]. It takes into account an additional effect of memory by updating both ω_A and ω_B continually. The *eligibility trace* can be interpreted as a description of how psychological perception of information refreshes or decays in response to whether or not an external stimulus is reinforced. The eligibility traces (as presented in [11]) denoted by $\phi_A(t)$ and $\phi_B(t)$ for choices A and B respectively, evolve according to

$$\phi_{Z(t)}(t+1) = 1 + \phi_{Z(t)}(t)e^{-\frac{1}{\tau}} \tag{3.13}$$

$$\phi_{\bar{Z}(t)}(t+1) = \phi_{\bar{Z}(t)}(t)e^{-\frac{1}{\tau}} \tag{3.14}$$

with initial values $\phi_A(0) = \phi_B(0)$, where $\tau > 0$ is a parameter that determines the decaying effects of memories.

With the eligibility traces included, ω_A and ω_B are updated according to

$$\omega_A(t+1) = \omega_A(t) + \lambda[r(t) - \omega_{Z(t)}(t)]\phi_A(t) \quad (3.15)$$

$$\omega_B(t+1) = \omega_B(t) + \lambda[r(t) - \omega_{Z(t)}(t)]\phi_B(t) \quad (3.16)$$

where the eligibility traces ϕ_A and ϕ_B act as time-varying weighting factors. When τ is small, the update rule in the eligibility trace model (3.15) and (3.16) reduces to that in the standard model (3.8) and (3.9).

Discretized Eligibility Trace

To analyze the impact of the dynamics of the eligibility traces on the evolution of ω_A and ω_B, we discretize the eligibility traces ϕ_A and ϕ_B. We set the learning rate λ in (3.15) and (3.16) to be its maximum value ($\lambda = 1$) which corresponds to the current reward having the strongest possible influence on the subject. Then

$$\omega_A(t+1) = \omega_A(t) + [r(t) - \omega_{Z(t)}(t)]\phi_A(t) \quad (3.17)$$

$$\omega_B(t+1) = \omega_B(t) + [r(t) - \omega_{Z(t)}(t)]\phi_B(t). \quad (3.18)$$

We discretize ϕ_A and ϕ_B as follows. For $Z \in \{A, B\}$, if Z is the most recent choice made, we set the value of ϕ_Z to be saturated at one because the impact of the current reward has been accounted for by setting λ to be its maximum value. We let ϕ_Z decay to zero immediately once the opposite choice $\bar{Z}$ has been chosen consecutively. With the eligibility trace modeled in this way we consider a process in which memory fades quickly.

When a switch in choice is made (from Z at time $t-1$ to $\bar{Z}$ at time t), let the stored memory of the unchosen alternative, ϕ_Z, take its value to be a small, positive number $\epsilon \in (0, 1)$. This is a simple model of a subject "forgetting" the value of an

unchosen reward. Then ϕ_A and ϕ_B take values in $\{0, \epsilon, 1\}$ and evolve according to

$$\phi_Z(t) = \begin{cases} 1 & \text{if } Z = x_1(t) \\ \epsilon & \text{if } Z = \bar{x}_1(t) = x_1(t-1) \\ 0 & \text{if } Z = \bar{x}_1(t) = \bar{x}_1(t-1) \end{cases} \quad t = 1, 2, 3, \ldots \tag{3.19}$$

The resulting model, defined by (3.12) and (3.17)-(3.18), is a deterministic limit of the soft-max choice model with discretized eligibility traces. Like the WSLS model, the deterministic limit (as presented here) can replicate some human behavior. We note, however, that it does not accurately represent the majority of human behavior since it lacks stochasticity.

3.4.4 Soft-Max Choice Model with Social Feedback

Each individual in a group of decision makers can be modeled with a soft-max choice model as described above in Section 3.4.2. Social feedback is introduced with a feedback term that interconnects the decision makers. Here, we model decision making for one focal individual who receives feedback on the choices of M other decision makers, none of whom receives any feedback themselves. In effect, we consider a team in a network with a directed interconnection graph. Doing so makes the model tractable in a way that allows for developing a Markov chain to represent the evolving state of the model in the TAFC task. We are able to relax the directed assumption to perform a numerical analysis of an undirected network also. Each of these analyses appears in Chapter 5.

The approach developed by Nedic et al. [56, 57] for deriving a choice feedback law is to bias anticipated rewards with a feedback parameter ν that reinforces the focal individual's tendency to choose A (respectively, B) when a majority of the M others choose A (respectively, B). With this feedback, the probability that the focal

individual chooses A in the next time step is

$$p_A(t+1,\nu) = \frac{1}{1+e^{-\mu(w_A(t)-w_B(t)+\nu u(t))}} \tag{3.20}$$

$$u(t) = \begin{cases} 1 & \text{if } |A| \geq \lceil \frac{M+1}{2} \rceil \\ -1 & \text{if } |B| \geq \lceil \frac{M+1}{2} \rceil \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

where $|A|$ refers to the number of others (not receiving feedback) who choose A at time t, and $\lceil \cdot \rceil$ gives the smallest integer greater than or equal to its argument. Again, w_A and w_B represent the decision maker's learned anticipated reward for choice A and B. The anticipated rewards are also modeled by (3.8) - (3.9) so that

$$\begin{aligned} w_Z(t+1) &= (1-\lambda)\omega_Z(t) + \lambda r(t) \\ w_{\bar{Z}}(t+1) &= w_{\bar{Z}}(t) \qquad t = 0, 1, 2, \ldots \end{aligned}$$

where $\bar{\cdot}$ denotes the "not" operator, Z corresponds to the choice at time t, and $\bar{Z}$ the choice not made at time t. The parameter λ acts as a learning rate and satisfies $\lambda \in [0, 1]$. Larger λ also implies less "memory" since the influence of previous rewards decays. Note that the no-feedback case (3.7) is equivalent to $p_A(t+1, 0)$ in (3.20).

Several different choice feedback models were presented in [56, 57]. Models with different numbers of fitting parameters were compared using the Akaike information criterion together with estimated maximum likelihoods for the prediction of choice sequences. The choice feedback model (3.20)-(3.21) performed well in those tests. Another model that uses an additional "contrarion" parameter is shown in [57] to fit the data more successfully. This additional parameter, which represents a decision maker's tendency to agree or disagree with the M other decision makers, has a

history dependency that breaks the Markov property. We find the model defined in this section particularly convenient since in Chapter 5 we can make two reasonable assumptions to develop a Markov chain and corresponding analysis of the evolution of $y(t)$ in the TAFC task.

3.5 Mixed Teams

By studying the TAFC task with social feedback from Section 3.3, we develop a framework and formal understanding of decision making in teams. The analysis of Chapter 5, which pertains to the soft-max choice model with social feedback, defined in Section 3.4.4, yields results that can be applied to human and also *designed* decision makers. By designing some of the decision makers in a team to be automated and make choices using common models, we create *mixed teams* whose dynamics are relatively well-understood. This allows us to apply our tools for analysis and prediction, thereby ensuring that mixed teams of decision makers will exhibit desirable behavior, or perform more optimally, than humans would on their own.

We consider each member of the team to have the same role, that they must chose between A or B, sequentially in time, but they receive additional information, and also share information of their own. Decision makers that we consider may be human, or may be automated. New experiments explore our ability to design decision makers and influence performance of a team. In Chaper 7, we design automated decision-making and corresponding social feedback provided to a human subject in an effort to determine the influence of the feedback and network interconnection properties.

Chapter 4

Convergence in Deterministic Decision-Making Models

While it is the case that choice sequences made by humans are not exemplary of a deterministic choice rule, we show in this chapter that deterministic models are capable of reproducing similar behavior, and even replicating convergent behavior as seen in data. In this chapter, we analyze the deterministic decision-making models described in Sections 3.4.1 and 3.4.3 for human behavior in the TAFC task without social feedback. We prove conditions for each of the decision-making models to converge to the *matching behavior* exhibited extensively in data [57, 41, 56]. See Section 3.2.4 for a description of matching behavior and a matching point in reward structures for the TAFC task.

The MS reward structure shown in Figure 3.1(a) is the simplest reward structure with a matching point. Both the RO and CG structure of Figure 3.1 also have a matching point. The convergence results of this chapter apply locally to this point in these and other reward structures. Attraction to the matching point in these more complex reward structures is an important factor in the context of social decision

making; attraction to the matching point may be in conflict with social influences. These issues in the social context are considered in Chapters 6 and 7.

The goal of the work here is to develop a framework and begin our analysis with two of the deterministic decision-making models that are relevant for the TAFC task.

In Section 4.2, we prove convergence to matching for the simple WSLS model introduced in Section 3.4.1. In Section 4.3 we prove convergence to matching for the deterministic limit of the soft-max choice model (as defined in Section 3.4.3).

Our proof of convergence to a neighborhood of the matching point was first published for the WSLS decision-making model in [18, 19]. A related analysis for the WSLS model is performed in [87]. Our proof of convergence for the deterministic limit of the choice model was first published in [19]. The theory in this Chapter was developed in collaboration with Ming Cao.

4.1 Convergence to Matching Points

Decision makers in the TAFC, when faced with reward structures that contain matching points, typically make choices in a way that drives y, their proportion of A in the recent length-N choice history, toward the value at the matching point. This is a well-known phenomenon in human behavioral experiments [40, 41]: human decision makers in TAFC tasks converge to choice sequences in the neighborhood of the matching point for a variety of reward structures. However, there are relatively few results that prove conditions for this phenomenon given well-established models like the soft-max choice model. In [53], Montague and Berns argue based on an assumption (see also Assumption 2(a) in Section 5.1) that the soft-max choice model should converge to matching behavior in the TAFC task.

Denote by y^* the value of y at the matching point, i.e. at the intersection of the two curves r_A and r_B. We consider the generic case when

$$y^* \notin \mathcal{Y}, \tag{4.1}$$

i.e., y^* is not an integer multiple of $1/N$. In the non-generic case, when $y^* \in \mathcal{Y}$, a tighter convergence result applies. Let y^l denote the greatest element in $\mathcal{Y}$ that is smaller than y^* and let y^u denote the smallest element in $\mathcal{Y}$ that is greater than y^*. Let $y^{l'} = y^l - 1/N$ and $y^{u'} = y^u + 1/N$. Define

$$\mathcal{L} \triangleq [y^l, y^u] \qquad \text{and} \qquad \mathcal{L}' \triangleq [y^{l'}, y^{u'}].$$

So that $\mathcal{L}'$ is well defined, let $1/N < y^* < (N-1)/N$ and $N \geq 3$.

4.2 Convergence of the WSLS Model

In this section, we analyze the convergence behavior of the WSLS system (3.1)-(3.6). We consider the set of reward curves such that r_A decreases monotonically and r_B increases monotonically with increasing y, i.e.,

$$\frac{d}{dy} r_A(y) < 0, \quad \frac{d}{dy} r_B(y) > 0, \quad \forall y \in [0,1]\ . \tag{4.2}$$

The set of reward curves defined by (4.2) includes the linear MS curves of Figure 3.1(a) as well as a more general class of nonlinear reward curves. It also includes the RO reward curves of Figure 3.1(b) locally about the matching point, as well as the CG of Figure 3.1(c).

The linear curves used in the experiments [53] satisfy the conditions (4.1), (4.2) and (4.3), so the analysis in this section provides an analytical understanding of

human decision-making dynamics in two-alternative forced-choice tasks of the same type. We prove both a local and a global convergence result to the matching point for the WSLS.

To avoid considering a class of limit cycles that are not thought to be relevant to human decision-making, it is necessary to consider the WSLS model in tasks that have reward structures satisfying

$$\frac{1}{3} \leq y^* \leq \frac{2}{3}. \tag{4.3}$$

4.2.1 Local convergence

The WSLS decision maker exhibits oscillatory behavior when $y(t)$ is near y^*. In the following theorem we prove that if $y \in \mathcal{L}$, i.e., the decision trajectory gets near the matching point y^*, then $y \in \mathcal{L}'$ for all future time. This means that the trajectory remains closed.

Theorem 1. *For system (3.1)-(3.6) satisfying conditions (4.1)-(4.3), if $y(t_1) \in \mathcal{L}$ for some $t_1 > 0$, then $y(t) \in \mathcal{L}'$ for all $t \geq t_1$.*

Theorem 1 is best understood by examining a typical trajectory in the task. A proof of Theorem 1 follows, but in order to introduce the notation used in this chapter, consider Figure 4.1. We define $p_1 = (y^l, r_A(y^l))$, $p_2 = (y^l, r_B(y^l))$, $p_3 = (y^u, r_A(y^u))$ and $p_4 = (y^u, r_B(y^u))$. Figure 4.1 shows these points for an example set of reward curves.

Consider a trajectory that starts at time $t = t_1$ with $y(t_1) \in \mathcal{L}$ with the MS reward structure shown in Figure 4.1 as an example. For illustration, suppose we are given a set of initial conditions $y(t_1) = y^u$, $x_1(t_1) = A$, $x_N(t_1) = B$ and suppose $x_1(t_1 + 1) = B$. Then $y(t_1 + 1) = y(t_1) = y^u$ and the reward $r(t_1 + 1) = r_B(y^u) > r_A(y^u) = r(t_1)$. In view of (3.6), we know that $x_1(t_1 + 2) = B$. If $x_N(t_1 + 1) = A$,

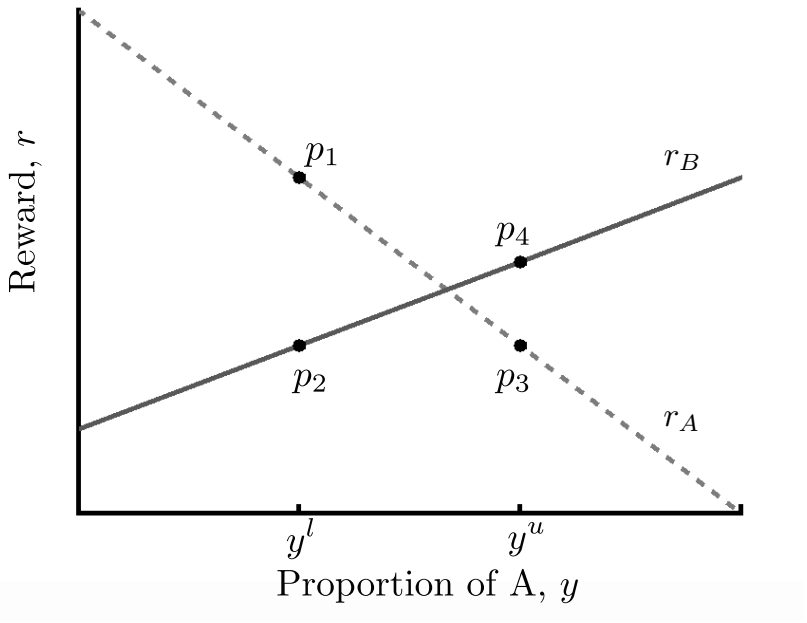


Figure 4.1: Points p_1, p_2, p_3, and p_4 used to examine trajectories around the matching point.

then $y(t_1+2) = y(t_1+1) - 1/N = y^l$ and $r(t_1+2) = r_B(y^l) < r_B(y^u) = r(t_1+1)$. Again by (3.6), it must be true that $x_1(t_1+3) = A$. Suppose $x_N(t+2) = B$, then $y(t_1+3) = y^u$.

Tracking the system's trajectory trajectory in this way shows how the decision maker moves from p_3, to p_4, to p_2 and back to p_3 in Figure 4.1. One may arrive at the conclusion that once $y(t)$ enters $\mathcal{L}$, it will stay in $\mathcal{L}$. This is not, however, necessarily true.

Consider a counterexample in which a trajectory again starts at p_3. However, let $x_N(t_1) = B$ and $x_1(t_1+1) = A$. Then $y(t_1+2) = y^u + 1/N \notin \mathcal{L}$. Although $\mathcal{L}$ is not an invariant set for $y(t)$, trajectories of $y(t)$ starting in $\mathcal{L}$ will always remain in $\mathcal{L}'$. This example motivates Theorem 1. In the following we prove Theorem 1

Proof

To prove Theorem 1, we first prove the following four lemmas.

Lemma 1. *For system (3.1)-(3.6), satisfying conditions (4.1)-(4.3), if* $x_1(t_1) = A$, $x_1(t_1 + 1) = A$ *and* $y(t_1) < 1$ *for some* $t_1 \geq 0$, *then there exists* $0 \leq \tau \leq N$ *such that* $y(t) = y(t_1)$ *for* $t_1 \leq t \leq t_1 + \tau$ *and* $y(t_1 + \tau + 1) = y(t_1) + 1/N$.

Proof of Lemma 1: If $x_N(t_1) = B$, then $y(t_1+1) = y(t_1)+1/N$. So the conclusion holds for $\tau = 0$. On the other hand, if $x_N(t_1) = A$, then $y(t_1 + 1) = y(t_1)$ and $r(t_1 + 1) = r_A(y(t_1 + 1)) = r_A(y(t_1)) = r(t_1)$. According to (3.6), $x_1(t_1 + 2) = A$. In fact A will be repeatedly chosen as long as the value of x_N remains A. However, since $y(t_1) < 1$, there must exist $0 \leq \tau < N$ such that $x_N(t) = A$ for $t_1 \leq t \leq t_1 + \tau$ and $x_N(t_1 + \tau + 1) = B$. Accordingly, the conclusion holds. □

One can prove the following lemma, the counterpart to Lemma 1, with a similar argument.

Lemma 2. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $x_1(t_1) = B$, $x_1(t_1 + 1) = B$ *and* $y(t_1) > 0$ *for some* $t_1 \geq 0$, *then there exists* $0 \leq \tau \leq N$ *such that* $y(t) = y(t_1)$ *for* $t_1 \leq t \leq t_1 + \tau$ *and* $y(t_1 + \tau + 1) = y(t_1) - 1/N$.

Now we further study behavior of the system when its trajectory starts on the left of the matching point y^*.

Lemma 3. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $y(t_1) < y^*$ *and* $y(t_1 + 1) = y(t_1) - 1/N > 0$ *for some* $t_1 \geq 0$, *then there exists* $0 \leq \tau \leq N$ *such that*

$$y(t) = y(t_1) - 1/N \quad \textit{for } t_1 \leq t \leq t_1 + \tau \tag{4.4}$$

and

$$y(t_1 + \tau + 1) = y(t_1). \tag{4.5}$$

Proof of Lemma 3: We find it convenient to prove this lemma by labeling the following four points: $s_1 = (y(t_1), r_A(y(t_1)))$, $s_2 = (y(t_1), r_B(y(t_1)))$, $s_3 = (y(t_1) - 1/N, r_A(y(t_1) - 1/N))$, $s_4 = (y(t_1) - 1/N, r_B(y(t_1) - 1/N))$, as shown

in Figure 4.2.

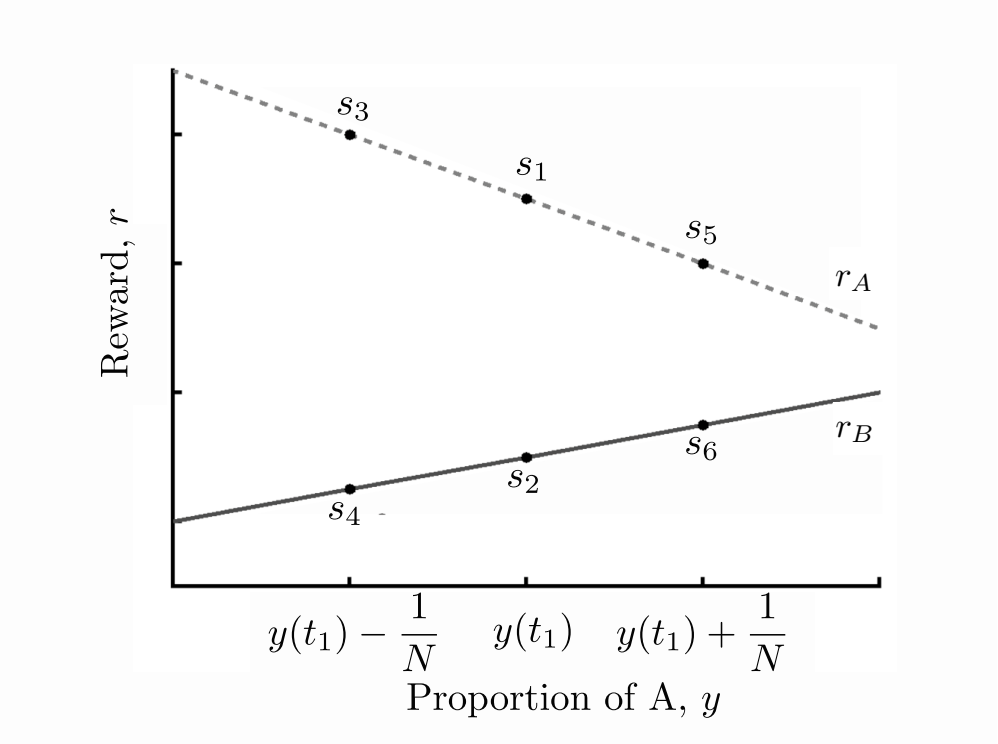


Figure 4.2: Points s_1, s_2, s_3, s_4, s_5, and s_6 used in the proofs of Lemma 3 and Lemma 7.

We denote the reward values at these four points by $r|_{s_i}$, $i = 1, \ldots, 4$. Then $r(t_1) = r|_{s_1}$ or $r|_{s_2}$. Since $y(t_1 + 1) < y(t_1)$, it must be true that $x_1(t_1 + 1) = B$, then $r(t_1 + 1) = r|_{s_4}$. Since $r|_{s_4} < r|_{s_2} < r|_{s_1}$, we know $x_1(t_1 + 2) = A$. So at $t_1 + 2$, the system trajectory moves from s_4 to either s_1 or s_3. If the former is true, the conclusion holds for $\tau = 2$. If the latter is true, since $r|_{s_3} > r|_{s_4}$, it follows that $x_1(t_1 + 3) = A$. By applying Lemma 1, we know (4.4) and (4.5) hold. □

Similarly, we consider the situation when $y(t_1) > y^*$ and $y(t_1+1) = y(t_1)+1/N < 1$ for some $t_1 \geq 0$. Denote four points: $r_1 = (y(t_1), r_A(y(t_1)))$, $r_2 = (y(t_1), r_B(y(t_1)))$, $r_3 = (y(t_1) + 1/N, r_A(y(t_1) + 1/N))$ and $r_4 = (y(t_1) + 1/N, r_B(y(t_1) + 1/N))$. Using the fact that $r|_{r_3} < r|_{r_1} < r|_{r_2}$ and a similar argument as that in the proof of Lemma 3, we can prove the following result.

Lemma 4. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $y(t_1) > y^*$ *and* $y(t_1 + 1) = y(t_1) + 1/N$ *for some* $t_1 \geq 0$*, then there exists* $0 \leq \tau \leq N$ *such that*

$$y(t) = y(t_1) + 1/N \quad \textit{for } t_1 \leq t \leq t_1 + \tau \tag{4.6}$$

and

$$y(t_1 + \tau + 1) = y(t_1). \tag{4.7}$$

Now we are in a position to prove Theorem 1.

Proof of Theorem 1: If $y(t) \in \mathcal{L}$ for all $t \geq t_1$, then the conclusion holds trivially. Now suppose this is not true. Let $t_2 > t_1$ be the first time for which $y(t) \notin \mathcal{L}$. Then it suffices to prove the claim that the trajectory of $y(t)$ starting at $y(t_2)$ stays at $y(t_2)$ for a finite time and then enters $\mathcal{L}$. Note that $y(t_2)$ equals either $y^l - 1/N$ or $y^u + 1/N$. Suppose $y(t_2) = y^l - 1/N$, then the claim follows directly from Lemma 3; if on the other hand, $y(t_2) = y^u + 1/N$, then the claim follows directly from Lemma 4. □

The local result is applicable to reward structures that have a local matching point and satisfy (4.2) for y in a neighborhood of y^*. This includes the rising optimum reward curves of Figure 3.1(a). This analysis has uncovered limit cycles about points other than the matching point. The convergence results of this section apply for general $N \geq 6$. For lower values of N the system degenerates and the output y may converge to 0 or 1.

4.2.2 Global convergence

Theorem 1 is a local convergence result which applies in the neighborhood $\mathcal{L}$ of the matching point y^*. It can be further shown that the convergence is global. In Theorem 2 we prove that for most initial conditions, the decision trajectory will converge to $y \in \mathcal{L}'$, i.e., it will converge to this neighborhood of the matching point.

It is easy to check that if the system starts with the initial condition $y(0) = 0$ and $x_1(1) = B$ or the initial condition $y(0) = 1$ and $x_1(1) = A$, then the trajectory of $y(t)$ will stay at its initial location. It will also be shown that when $y^* < \frac{1}{3}$ or $y^* > \frac{2}{3}$, a limit cycle of period three not containing y^* may appear. Thus it is necessary that condition (4.3) be satisfied, i.e., $\frac{1}{3} \leq y^* \leq \frac{2}{3}$.

To prove Theorem 2, we first show in Proposition 1 that if the trajectory $y(t)$ starts in $(0, 1)$ then the trajectory always enters $\mathcal{L}$ after a finite time. Proposition 1 together with Theorem 1 then prove Theorem 2.

Proposition 1. *For system (3.1)-(3.6), satisfying conditions conditions (4.1)-(4.3), for any initial condition satisfying* $0 < y(0) < 1$*, there is a finite time* $T > 0$ *such that* $y(T) \in \mathcal{L}$*.*

To prove Proposition 1, we need to prove the following four lemmas.

Lemma 5. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $y(t_1) < y^*$*,* $y(t_1 + 1) = y(t_1)$ *and* $x_1(t_1 + 1) \neq x_1(t_1)$ *for some* $t_1 \geq 0$*, then there exists a finite* $\tau > 0$ *such that*

$$y(t_1 + \tau) = y(t_1) + 1/N. \tag{4.8}$$

Proof of Lemma 5: There are two cases to consider. (a) Suppose $x_1(t_1 + 1) = A$ and $x_1(t_1) = B$. Since $y(t_1 + 1) = y(t_1) < y^*$, we know $r(t_1 + 1) = r_A(y(t_1 + 1)) = r_A(y(t_1)) > r_B(y(t_1)) = r(t_1)$, so $x_1(t_1 + 2) = A$. Then the conclusion follows from Lemma 1. (b) Now suppose instead $x_1(t_1+1) = B$ and $x_1(t_1) = A$. Again since $y(t_1+1) = y(t_1) < y^*$, we know $r(t_1 + 1) = r_B(y(t_1 + 1)) = r_B(y(t_1)) < r_A(y(t_1)) = r(t_1)$, so $x_1(t_1 + 2) = A$. As a result, either $y(t_1 + 2) = y(t_1) + 1/N$ or $y(t_1 + 2) = y(t_1 + 1)$. If the former is true, then the conclusion holds for $\tau = 2$; if the latter is true, then the discussion reduces to that in (a). □

Using a similar argument, one can prove the following lemma, which is the counterpart to Lemma 5.

Lemma 6. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $y(t_1) > y^*$, $y(t_1+1) = y(t_1)$ *and* $x_1(t_1+1) \neq x_1(t_1)$ *for some* $t_1 \geq 0$, *then there exists a finite* $\tau > 0$ *such that*

$$y(t_1+\tau) = y(t_1) - 1/N. \tag{4.9}$$

Now we show that the system will approach the matching point.

Lemma 7. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $0 < y(t_1) < y^l$ *and* $y(t_1+1) = y(t_1) - 1/N$ *for some* $t_1 \geq 0$, *then there exists a finite* $\tau > 0$ *such that*

$$y(t_1+\tau) = y(t_1) + 1/N. \tag{4.10}$$

Proof of Lemma 7: Denote the six points $s_1 = (y(t_1), r_A(y(t_1)))$, $s_2 = (y(t_1), r_B(y(t_1)))$, $s_3 = (y(t_1) - 1/N, r_A(y(t_1) - 1/N))$, $s_4 = (y(t_1) - 1/N, r_B(y(t_1) - 1/N))$, $s_5 = (y(t_1) + 1/N, r_A(y(t_1) + 1/N))$ and $s_6 = (y(t_1) + 1/N, r_B(y(t_1) + 1/N))$, as shown in Figure 4.2. Since $y(t_1+1) < y(t_1)$, it must be true that $x_1(t_1+1) = B$.

If $x_1(t_1) = A$, we know from t_1 to t_1+1, the system trajectory moves from s_1 to s_4. Since $r(t_1+1) = r|_{s_4} < r|_{s_1} = r(t_1)$, we know $x_1(t_1+2) = A$. Then at t_1+2, the trajectory moves to either s_1 or s_3. We discuss these two cases separately.

- (a) If at t_1+2 the trajectory moves to s_1, since $r|_{s_1} > r|_{s_4}$, it follows that $x_1(t_1+3) = A$. In view of Lemma 1, the conclusion holds.

- (b) If at t_1+2, the trajectory moves to s_3, since $r|_{s_3} > r|_{s_4}$, we know $x_1(t_1+3) = A$.

Thus, in case (b), from Lemma 1, there exists a finite time $t_2 < N$ at which the trajectory moves from s_3 to s_1. Because $r|_{s_1} < r|_{s_3}$, we have $x_1(t_2+1) = B$. Then at time t_2+1, the trajectory moves to either s_2 or s_4.

Now consider two sub-cases: $b(1)$ Suppose the former is true, that the trajectory goes to s_2. The conclusion follows directly from Lemma 5. $b(2)$ Suppose the latter is true, that the trajectory goes to s_4. Because $r|_{s_4} < r|_{s_1}$, $x_1(t_2+2) = A$. Then $y(t)$ will remain strictly less than $y(t_1) + 1/N$ if a cycle of $s_4 \to s_3 \to s_1 \to s_4$ is formed. In fact, from the analysis above, this is the only potential scenario in case $b(2)$ where $y(t) < y(t_1) + 1/N$ for all $t \geq t_1$. Were such a cycle to appear, A would be chosen at least twice as often as B. However, because $y(t_1) < y^l = y^* - 1/N < \frac{1}{3}$, it must be true that the proportion of A in $x_i(t_1)$, $1 \leq i \leq N$, is less than $\frac{1}{3}$. Thus such a cycle can never happen. So the conclusion also holds for the sub-case $b(2)$.

If on the other hand, $x_1(t_1) = B$, we know from t_1 to $t_1 + 1$, the system trajectory moves from s_2 to s_4. Since $r|_{s_4} < r|_{s_2}$, we know $x_1(t_1+2) = A$. So at $t_1 + 2$, the trajectory moves to s_3 or s_1. If the former is true, the discussion reduces to ruling out the possibility of forming a cycle of $s_4 \to s_3 \to s_1 \to s_4$ which we have done in $b(2)$. Otherwise, if the latter is true, since $r|_{s_1} > r|_{s_4}$, we know $x_1(t_1+3) = A$. From Lemma 1 we know there exists a finite time t_3 at which $y(t_3) = y(t_1) + 1/N$, and thus the conclusion holds for $\tau = t_3 - t_1$. Through consideration of each of the above arguments, we have proved Lemma 7. □

Lemma 8 is the counterpart of Lemma 7 in that it applies in the same way only on the opposite side of the matching point. Using Lemmas 2, 6 and a similar argument as in the proof of Lemma 7, one can prove the following lemma.

Lemma 8. *For system (3.1)-(3.6), with conditions (4.1)-(4.3) satisfied, if* $y^u < y(t_1) < 1$ *and* $y(t_1+1) = y(t_1) + 1/N$ *for some* $t_1 \geq 0$*, then there exists a finite* $\tau > 0$ *such that*

$$y(t_1 + \tau) = y(t_1) - 1/N. \tag{4.11}$$

Now we are in a position to prove Proposition 1.

Proof of Proposition 1: For any $0 < y(0) < 1$, either $y(1) = y(0) + 1/N$, or

$y(1) = y(0)$, or $y(1) = y(0) - 1/N$. We will discuss these three possibilities in each of two cases. First consider the case where $y(0) < y^l$. If $y(1) = y(0) - 1/N$, according to Lemma 7, there is a finite time t_1 for which $y(t_1) > y(0)$. If $y(1) = y(0)$ and $x_1(1) \neq x_1(0)$, according to Lemma 5, there is a finite time t_2 for which $y(t_2) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = A$, according to Lemma 1, there is a finite time t_3 for which $y(t_3) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = B$, according to Lemma 2, there is a finite time $\bar{t}_4$ for which $y(\bar{t}_4 - 1) = y(0)$ and $y(\bar{t}_4) = y(\bar{t}_4 - 1) - 1/N$. Then according to Lemma 7, there is a finite time t_4 for which $y(t_4) > y(0)$. So for all possibilities of $y(1)$ there is always a finite time $\bar{t} \in \{1, t_1, t_2, t_3, t_4\}$ for which $y(\bar{t}) > y(0)$. Using this argument repeatedly, we know that there exists a finite time T_1 at which $y(T_1) = y^l \in \mathcal{L}$. Now consider the other case where $y(0) > y^u$, then using similar arguments, one can check that there exists a finite time T_2 for which $y(T_2) = y^u \in \mathcal{L}$. Hence, we have proven the existence of T which lies in the set $\{T_1, T_2\}$. □

Combining the conclusions in Theorem 1 and Proposition 1, we have proven Theorem 2, which describes the global convergence property of $y(t)$.

Theorem 2. *For any initial condition of the system (3.1)-(3.6) satisfying* $0 < y(0) < 1$ *with conditions (4.1)-(4.3) satisfied, there exists a finite time* $T > 0$ *such that for any* $t \geq T$, $y(t) \in \mathcal{L}'$.

4.3 Convergence of the Deterministic Limit of the Soft-Max Choice Model

The deterministic limit of the soft-max choice model (derived from the stochastic choice model by taking μ to infinity) exhibits very similar behavior with respect to convergence in tasks with reward structures containing at least one matching point. In this section we are able to prove convergence of $y(t)$ to $\mathcal{L}'$ (Theorem 3) for the

deterministic limit of the model with a discretized eligibility trace and reward structures with matching points. As in the analysis for the WSLS model, we consider the general case (4.1) when $y^* \notin \mathcal{Y}$. Also like the analysis for the WSLS model, these results generalize to nonlinear curves. For clarity of presentation, however, we specialize to intersecting linear reward curves defined by

$$\begin{aligned} r_A(y) &= k_A y + c_A \\ r_B(y) &= k_B y + c_B \end{aligned} \tag{4.12}$$

where $k_A < 0$, $k_B > 0$ and $c_A, c_B > 0$.

We first look at the case when the subject does not switch choice at a given time t_0.

Lemma 9. *For any $t_0 > 0$, if $y(t_0 - 1) < 1$ and $x_1(t_0 - 1) = x_1(t_0) = A$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = A$ for all $t_0 \leq t \leq t_1$ and $y(t_1) = y(t_0-1)+1/N$.*

Proof of Lemma 9: If $x_N(t_0 - 1) = B$, then $y(t_0) = y(t_0 - 1) + 1/N$ and so the conclusion holds for $t_1 = t_0$. If on the other hand $x_N(t_0-1) = A$, then $y(t_0) = y(t_0-1)$. From (3.19) we know that $\phi_A(t_0 - 1) = \phi_A(t_0) = 1$ and $\phi_B(t_0) = 0$. Then it follows from (3.17) that $w_A(t_0 + 1) = r(t_0) = r_A(y(t_0)) = r_A(y(t_0 - 1)) = w_A(t_0)$ and from (3.18) that $w_B(t_0 + 1) = w_B(t_0)$. Since $x_1(t_0 - 1) = x_1(t_0) = A$, from (3.12) it must be true that $w_A(t_0) > w_B(t_0)$. Thus we know $w_A(t_0 + 1) > w_B(t_0 + 1)$, so again from (3.12), we have $x_1(t_0 + 1) = A$. In fact, the choice of A will be repeatedly chosen as long as the value of x_N remains A. However, since $y(t_0 - 1) < 1$, there must exist $t_1 \leq t_0 + N$ such that $x_N(t_1 - 1) = B$ and then for the same t_1, we have $x_1(t) = A$ for all $t_0 \leq t \leq t_1$, and $y(t_1) = y(t_0 - 1) + 1/N$. □

The following lemma is the counterpart to Lemma 9. What differs is only that Lemma 10 applies to the opposite side of the reward structure. Therefore, a similar argument proves Lemma 10.

Lemma 10. *For any $t_0 \geq 0$, if $y(t_0 - 1) > 0$ and $x_1(t_0 - 1) = x_1(t_0) = B$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = B$ for all $t_0 \leq t \leq t_1$, and $y(t_1) = y(t_0 - 1) - 1/N$.*

Lemmas 9 and 10 imply that if $Z \in \{A, B\}$ is repeatedly chosen, then the anticipated reward for choice Z decreases as a result of the change in y while the anticipated reward for the alternative $\bar{Z}$ stays the same because the eligibility trace $\phi_{\bar{Z}}$ remains zero. Hence, a switch of choices must happen after a finite time. Now we look at the case when the subject switches choice at time $t_0 > 0$, namely $x_1(t_0) = \bar{x}_1(t_0 - 1)$. Then from (3.19), we have $\phi_{x_1(t_0-1)}(t_0) = \epsilon$; correspondingly from update rules (3.17) and (3.18), we have $w_{x_1(t_0-1)}(t_0 + 1) = w_{x_1(t_0-1)}(t_0) + \epsilon(r(t_0) - w_{\bar{x}_1(t_0-1)}(t_0)) = w_{x_1(t_0-1)}(t_0) + \epsilon(w_{\bar{x}_1(t_0-1)}(t_0 + 1) - w_{\bar{x}_1(t_0-1)}(t_0))$. Hence, the magnitude of ϵ is critical in updating the value of the anticipated reward when a switch of choices happens. It should be pointed out that in Section 3.4.3, to be consistent with the exponential decay rate for eligibility trace in [11], we have made an assumption that ϵ is a small number. This assumption can be stated by restricting the upper bound for the convex combination of points on the r_A and r_B lines.

Assumption 4. (Restricted Convex Combination) *For $y \in \mathcal{Y}$,*

$$(1 - \epsilon) \min\{r_A(y), r_B(y)\} + \epsilon \max\{r_A(y), r_B(y)\} < r_A(y^*) = r_B(y^*).$$

The following result guarantees that under specific circumstances the deterministic limit of the soft-max choice model will not get stuck in an oscillatory cycle such that recurring switches occur.

Lemma 11. *Suppose Assumption 4 is satisfied. For any $t_0 > 0$, if $x_1(t_0) = A$, $x_1(t_0 - 1) = B$ and $y(t_0 - 1) < y^*$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = A$ for all $t_0 \leq t \leq t_1$, and $y(t_1) > y^*$.*

Proof of Lemma 11: From (3.19) we know that $\phi_A(t_0) = 1$ and $\phi_B(t_0) = \epsilon$. So from (3.17) and (3.18), it follows that

$$w_B(t_0) = r(t_0 - 1) = r_B(y(t_0 - 1)), \tag{4.13}$$

$$w_A(t_0 + 1) = r(t_0) = r_A(y(t_0)), \tag{4.14}$$

and

$$w_B(t_0 + 1) = w_B(t_0) + \epsilon(r(t_0) - w_A(t_0)) = w_B(t_0) + \epsilon(r_A(y(t_0)) - w_A(t_0)). \tag{4.15}$$

Since $x_1(t_0) = A$, from (3.12) it must be true that $w_A(t_0) \geq w_B(t_0)$. Combining with (4.15), we have

$$w_B(t_0 + 1) \leq w_B(t_0) + \epsilon(r_A(y(t_0)) - w_B(t_0)) = (1 - \epsilon)w_B(t_0) + \epsilon r_A(y(t_0)).$$

Substituting (4.13), we have

$$w_B(t_0 + 1) \leq (1 - \epsilon)r_B(y(t_0 - 1)) + \epsilon r_A(y(t_0)).$$

Since $x_1(t_0) = A$ and $x_1(t_0 - 1) = B$, we know $y(t_0) = y(t_0 - 1)$ or $y(t_0) = y(t_0 - 1) + 1/N$. Since $y(t_0 - 1) < y^*$, it must be true that either $y(t_0) = y^u$ or $y(t_0) \leq y^l$. We consider these two cases separately. Case (a): $y(t_0) = y^u$. Set $t_1 = t_0$, then $y(t_1) > y^*$ holds trivially. Case (b): $y(t_0) \leq y^l$. Then $r_A(y^*) < r_A(y(t_0)) \leq r_A(y(t_0 - 1))$. Also,

$$w_B(t_0 + 1) \leq (1 - \epsilon)r_B(y(t_0 - 1)) + \epsilon r_A(y(t_0 - 1)) < r_A(y^*),$$

where the last inequality follows from Assumption 4. Combining these with (4.14) we know that

$$x_1(t_0+1) = A \tag{4.16}$$

and consequently $\phi_A(t_0+1) = 1$ and $\phi_B(t_0+1) = 0$. In fact, A will be repeatedly chosen, ϕ_A and ϕ_B will remain one and zero respectively until some finite time $t_1 > t_0$ for which $w_A(t_1) = r(t_1) = r_A(y(t_1))$ is less than or equal to $w_B(t_0+1)$ or $y(t_1) = 1$. Since $w_B(t_0+1) < r_A(y^*)$, it follows that $y(t_1) > y^*$. So we have proved the conclusion for case (b) and the proof is complete. □

The following lemma is the counterpart to Lemma 11. Its proof is nearly identical since there is no structural difference between choices A and B in the TAFC task. This is especially true for the MS structure we consider in this analysis.

Lemma 12. *Suppose Assumption 4 is satisfied. For any $t_0 > 0$, if $x_1(t_0) = B$, $x_1(t_0-1) = A$ and $y(t_0-1) > y^*$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = B$ for all $t_0 \leq t \leq t_1$, and $y(t_1) < y^*$.*

In order to estimate the increment of anticipated rewards, which is necessary for this proof, we rely on a lemma which is true under the following assumption. Both simulations and experiments have indicated that the deterministic choice model fits subjects' behavior only when ϵ is bounded away from zero. Hence, we make the following assumption:

Assumption 5. (Bounded ϵ) *The positive number ϵ is bounded below satisfying*

$$\epsilon \geq \max\left\{ \frac{-k_A/N}{r_A(y^{l'}) - r_B(y^{l'})}, \frac{k_B/N}{r_B(y^{u'}) - r_A(y^{u'})}, \frac{r_A(y^{u'}) - r_B(y^{l'})}{r_A(y^l) - r_B(y^l)}, \frac{r_B(y^{l'}) - r_A(y^{u'})}{r_B(y^u) - r_A(y^u)} \right\}.$$

Assumption 5 allows us the following lemma:

Lemma 13. *Suppose Assumption 5 is satisfied. For any $t_0 > 0$, if $y(t_0) < y^{l'}$, $x_1(t_0-1) = x_1(t_0) = B$ and $x_1(t_0+1) = A$, then $w_B(t_0+2) \geq w_B(t_0+1) - k_A/N$.*

Proof of Lemma 13: Since $x_1(t_0-1) = x_1(t_0) = B$ and $x_1(t_0+1) = A$, it follows that

$$w_B(t_0+1) \leq w_A(t_0+1) = w_A(t_0) < w_B(t_0) \tag{4.17}$$

where

$$w_B(t_0+1) = r_B(y(t_0)) < r_B(y(t_0-1)) = w_B(t_0). \tag{4.18}$$

Then

$$\begin{aligned} w_B(t_0+2) &= w_B(t_0+1) + \epsilon\Big(w_A(t_0+2) - w_A(t_0+1)\Big) \\ &\geq w_B(t_0+1) + \epsilon\Big(w_A(t_0+2) - w_B(t_0)\Big) \\ &= w_B(t_0+1) + \epsilon\Big(r_A(y(t_0+1)) - r_B(y(t_0-1))\Big). \end{aligned}$$

From $x_1(t_0-1) = x_1(t_0) = B$ and $x_1(t_0+1) = A$, we know that $y(t_0) = y(t_0-1)$ or $y(t_0) = y(t_0-1) - 1/N$ and $y(t_0+1) = y(t_0)$ or $y(t_0+1) = y(t_0) + 1/N$. Since $y(t_0) < y^{l'}$, it follows that $y(t_0+1) \leq y^{l'}$ and $y(t_0-1) \leq y^{l'}$. Because of the monotonicity of r_A and r_B, it follows that $w_B(t_0+2) \geq w_B(t_0+1) + \epsilon(r_A(y^{l'}) - r_B(y^{l'}))$. Using $\epsilon \geq \frac{-k_A/N}{r_A(y^{l'}) - r_B(y^{l'})}$ in Assumption 5, we reach the conclusion. □

Following similar steps and using $\epsilon \geq \frac{k_B/N}{r_B(y^{u'}) - r_A(y^{u'})}$ in Assumption 5, one can prove the following lemma which is the counterpart to Lemma 13.

Lemma 14. *Suppose Assumption 5 is satisfied. For any $t_0 > 0$, if $y(t_0) > y^{u'}$, $x_1(t_0-1) = x_1(t_0) = A$ and $x_1(t_0+1) = B$, then $w_A(t_0+2) \geq w_A(t_0+1) + k_B/N$.*

Decision epochs where subjects make a switch from one choice to another are critical. For this reason, we examine time instances within the set $\mathcal{T}$ which includes all instances for which $t > 0$ and $x_1(t) \neq x_1(t-1)$, i.e., $\mathcal{T} \stackrel{\Delta}{=} \{t : x_1(t) \neq x_2(t)\}$.
It is also necessary to consider some subsets of $\mathcal{T}$. We define $\mathcal{T}_A \stackrel{\Delta}{=} \{t : t \in \mathcal{T}, x_1(t) = A, y(t-1) < y^*\}$ and $\mathcal{T}_B \stackrel{\Delta}{=} \{t : t \in \mathcal{T}, x_1(t) = B, y(t-1) > y^*\}$.

As in the analysis of the WSLS model, we consider $x_i(0)$, $i = 1, \ldots, N$, where $0 < y(0) < 1$. Then $w_{x_1(0)}(1) = r(0) = r_{x_1(0)}(y(0))$ is determined correspondingly. To simplify the analysis and rule out degenerate cases, we make the following assumption about the value of $w_{\bar{x}_1(0)}(1)$.

Assumption 6. (Bounded $w_{\bar{x}_1(0)}(1)$) *Let* $0 < y(0) < 1$ *and* $w_{x_1(0)}(1) = r_{x_1(0)}(y(0))$. *The initial value* $w_{\bar{x}_1(0)}(1)$ *satisfies*

$$w_{\bar{x}_1(0)}(1) < w_{x_1(0)}(1), \tag{4.19}$$

$$w_{\bar{x}_1(0)}(1) < r_A(y^*), \tag{4.20}$$

and

$$w_{\bar{x}_1(0)}(1) > \max\{r_A(1), r_B(0), r_A(1) + (k_B + k_A)/N, r_B(0) - (k_B + k_A)/N\}. \tag{4.21}$$

Now we are ready to study how the deterministic choice model with the eligibility trace evolves with time.

Lemma 15. *Suppose all the Assumptions 4 - 6 are satisfied. Then,*

$$\mathcal{T} = \mathcal{T}_A \cup \mathcal{T}_B.$$

Proof of Lemma 15: From (4.19) we know that $x_1(1) = x_1(0)$ and in fact $x_1(0)$ will be repeatedly chosen until the reward for choosing $x_1(0)$ is below $w_{\bar{x}_1(0)}(1)$. Such a switch will always happen because of (4.21). Let t_0 denote the time at which such a switch happens, i.e., $x_1(t_0 - 1) = x_1(0)$ and $x_1(t_0) = \bar{x}_1(0)$. In view of (4.20), we know that $y(t_0 - 1) < y^*$ if $x_1(0) = B$ and $y(t_0 - 1) > y^*$ if $x_1(0) = A$. So $t_0 \in \mathcal{T}_A$ if $x_1(0) = B$ and $t_0 \in \mathcal{T}_B$ if $x_1(0) = A$, and thus $t_0 \in \mathcal{T}_A \cup \mathcal{T}_B$. By inspection, if $y(t_0 - 1) \in \mathcal{L}$, then either Lemma 9 or Lemma 10 is applicable to t_0; if on the other

hand, $y(t_0-1) \notin \mathcal{L}$, then either Lemma 11 or 12 is applicable to t_0. This implies that $x_1(t_0)$ will be repeatedly chosen such that the time of the next switch $t_1 \geq t_0$ satisfies $t_1 \in \mathcal{T}_{\bar{x}_1(t_0)} \subset \mathcal{T}_A \cup \mathcal{T}_B$. Further, Lemmas 9 - 12 can be applied again to t_1. Then, by induction, we know all time instances for which a switch happens belong to $\mathcal{T}_A \cup \mathcal{T}_B$. □

This analysis uncovers the fact that values of y^*, k_A and k_B affect the range of the interval containing y^* to which $y(t)$ converges. In this chapter, we are interested in the sufficient condition under which such an interval is $\mathcal{L}'$. We can write this explicitly and do so through the following assumption. Assumption 7 puts a condition on the relative value of critical points in the reward structure defined by r_A and r_B. These points correspond to values of y in $\mathcal{L}'$.

Assumption 7. (Points in $\mathcal{L}'$)

$$r_B(y^l) + \epsilon(r_A(y^u) - r_B(y^u)) \geq r_A(y^{u'}), \tag{4.22}$$

$$r_A(y^u) + \epsilon(r_B(y^l) - r_A(y^l)) \geq r_B(y^{l'}). \tag{4.23}$$

Proposition 2. *Suppose all the Assumptions 4-7 are satisfied. If $t_0 \in \mathcal{T}$, $y(t_0-1) \in \mathcal{L}'$, then $y(t) \in \mathcal{L}'$ for all $t \geq t_0$.*

Proof of Proposition 2: The conclusion can be proved by induction if we can prove the following fact: There is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$. In order to prove this fact, we need to consider four cases.

- Case (a): $x_1(t_0) = A$ and $y(t_0-1) = y^{l'}$. Then

  $$w_B(t_0+1) > r_B(y^{l'}) + \epsilon\bigg(r_A(y(t_0)) - r_B(y^l)\bigg) \geq r_B(y^{l'}) + \epsilon\bigg(r_A(y^l) - r_B(y^l)\bigg),$$

  where the first inequality holds since $r_B(y^l) > r_B(y^{l'})$ and the last inequality holds because $r_A(y(t_0)) \geq r_A(y^l)$. In view of the inequality $\epsilon \geq \frac{r_A(y^{u'}) - r_B(y^{l'})}{r_A(y^l) - r_B(y^l)}$ in

Assumption 5 and combining with Lemma 11, we know that there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_B$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.

- Case (b): $x_1(t_0) = A$ and $y(t_0 - 1) = y^l$. Then

$$w_B(t_0+1) \geq r_B(y^l) + \epsilon\bigg(r_A(y(t_0)) - w_A(t_0)\bigg) > r_B(y^l) + \epsilon\bigg(r_A(y^u) - r_B(y^u)\bigg),$$

  where the first inequality holds because $w_A(t_0) \geq w_B(t_0)$ and the last inequality holds because $r_A(y(t_0)) \geq r_A(y^u)$ and $w_A(t_0) = w_A(t_0 - 1) < w_B(t_0 - 1) = r_B(y^*) < r_B(y^u)$. Then in view of (4.22), we know $w_B(t_0 + 1) \geq r_A(y^{u'})$, and so there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_B$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.

- Case (c): $x_1(t_0) = B$ and $y(t_0 - 1) = y^u$. Following similar steps as in case (b) and using (4.23), we know there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_A$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.

- Case (d): $x_1(t_0) = B$ and $y(t_0 - 1) = y^{u'}$. Following similar steps as in (a) and using the inequality $\epsilon \geq \frac{r_B(y^{l'}) - r_A(y^{u'})}{r_B(y^u) - r_A(y^u)}$ in Assumption 5 and combining with Lemma 12, we know that there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_A$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.

In view of the discussion in cases (a)-(d), we conclude that the proof is complete. □

Proposition 3. *Suppose all the Assumptions 4-7 are satisfied. There is always a finite $T \in \mathcal{T}$ for which $y(T-1) \in \mathcal{L}'$.*

Proof of Proposition 3: From (4.21) in Assumption 6, we know that $x_1(0)$ cannot be repeatedly chosen for more than N times, and thus there is $t_0 \in \mathcal{T}_A \cup \mathcal{T}_B$. If $y(t_0 - 1) \in \mathcal{L}'$, then set $T = t_0$ and we reach the conclusion. If on the other hand, $y(t_0 - 1) \notin \mathcal{L}'$, then either Lemma 13 or 14 applies to $t_0 - 1$. Without loss of generality, suppose Lemma 13 applies, then $w_B(t_0 + 1) \geq w_B(t_0) - k_A/N$. By (4.21)

$w_B(t_0+1) > r_A(1)$. So, there exists a t_1 which is the smallest element in $\mathcal{T}_B$ such that $t_1 > t_0$. Then $w_A(t_1) \geq w_B(t_0+1) + k_A/N > w_B(t_0)$. In fact, one can check that switches will continue to exist and $w_{\bar{x}_1(t)}$ will be a monotonically strictly increasing function until $w_{\bar{x}_1(t)}$ reaches either $r_A(y^{u'})$ or $r_B(y^{l'})$ at some finite time T'. Set T to be the smallest element in $\mathcal{T}$ that is greater than T', then it must be true that $y(T-1) \in \mathcal{L}'$. □

Combining Propositions 2 and 3, we have proved the main result of this section.

Theorem 3. *For system(3.1)-(3.3), (3.12), (3.17)-(4.1) and (4.12), if Assumptions 4-7 are satisfied, then there exists a finite $T > 0$, such that $y(t) \in \mathcal{L}'$ for all $t \geq T$.*

This concludes the analysis of this chapter. We have proven conditions under which both the WSLS and deterministic choice model converge to matching points. In the following Chapter we open our analysis to consider the soft-max choice model with all of its stochasticity.

Chapter 5

Convergence for Independent Decision Makers with Stochastic Decision-Making Models

Since decision making in humans faced with complex tasks typically has random properties found in the choice sequences, it is desirable to model behavior with a stochastic decision-making strategy. In this Chapter we perform an analysis for a stochastic decision-making model fitted to data of human subjects in the TAFC task. Here we consider a single soft-max model decision maker faced with the TAFC task described in Section 3.2.2. In this Chapter we focus on the case where decision makers do not receive social feedback. In Chapter 6 we extend this analysis using the coupled decision-making model system of [56, 57] described in Section 3.4.2.

Making use of two assumptions defined in Section 5.1, we derive a distribution that gives the fraction of time at steady state that the decision maker spends with the proportion of A in their history corresponding to each value $y \in \mathcal{Y}$. The approach is to identify the task and decision making model as a Markov process in Section 5.2 and compute the equilibrium probability distribution for that process in Section 5.3.

We make use of the analytic description of the decision maker's long-term behavior in Section 5.4 to prove conditions under which the decision maker will converge to a matching point. We also derive explicit dependence of performance on the decision-making parameter μ, which quantifies the individual's tendency to explore. The same two assumptions used in this chapter will apply in the analysis of Chapter 6. At the end of this Chapter, in Section 5.5, we compare the analytic results to data from experiments with human subjects.

Our approach to analyze the soft-max choice model using a Markov process was first published with preliminary results in [78] and presented at the American Control Conference in Baltimore, MD in 2010. We also published this method of analysis, along with preliminary results for decision making with social feedback in [80]. The complete results of this chapter have been published in [79].

5.1 Assumptions

We make two assumptions for the task and soft-max choice model with the goal of developing a Markov chain to analyze the decision making behavior. We take these two assumptions to hold throughout this chapter and Chapter 6.

The state of the model decision maker in the TAFC task, described in Section 3.2.2, is the N-element decision history $x(t)$ and the two anticipated rewards $w_A(t)$ and $w_B(t)$. The assumptions made in this section reduce the state to the scalar variable $y(t)$; we show further in Section 5.2 that with these assumptions, the system dynamics can be described as a Markov process. While $y(t)$ is deterministically given by $x(t)$, since $x(t)$ evolves according to a stochastic process defined by the soft-max model, we take $y(t)$ as a random variable in our analysis in this chapter.

Assumption 8, as stated below, applies for all reward structures. We make Assumption 9(a) when considering the MS, CG and DG reward structures and Assumption 9(b) for the RO reward structure.

Assumption 8. $Pr\{x_k(t) = A|x(t)\} = y(t)$.

Assumption 9(a). $w_B(t) - w_A(t) = \Delta r(y(t))$

Assumption 9(b). $w_B(t) - w_A(t) = f(y(t))$, *where* $f(y)$ *is given by the curve in Figure 5.1.*

Assumption 8 implies that the yN A's and $(1-yN)$ B's in $x(t)$ are uniformly distributed in the finite history. This assumption is believed to hold for choice sequences when the decision making occurs over long time periods since for each $y(t)$ visited by the system, all possible combinations of ordering of $x(t)$ should also occur with approximately equal probability. Andrea Nedic has performed numerical simulations indicating this to be true on average.

Given this assumption, the state of the system can be represented by $y(t)$, $w_A(t)$ and $w_B(t)$. Assumption 9(a) or 9(b) allows further simplification by replacing dependence of the process on $w_A(t)$ and $w_B(t)$ by a single state, $y(t)$. Note that Δr used in Assumption 9(b) is defined by $\Delta r := r_B - r_A$.

Assumption 9(a) sets the difference in anticipated rewards at time t equal to the difference in rewards evaluated at $y(t)$. This assumption was introduced in a paper by Montague and Berns [53] who argue that it is true "on average". It is equivalent to stating that the decision maker has perfect knowledge of the rewards for each decision at time t. To further investigate the validity of this assumption, we performed a numerical study without using Assumption 9(a) by building a Markov chain with state $y(t)$, $w_A(t)$ and $w_B(t)$. We used $\lambda = 1$ as this agrees well with the fitted value of λ in the CG and DG tasks [57].

By computing an equilibrium distribution numerically, without Assumption 9(a), and comparing it to the equilibrium distribution we derive below in Section 5.3, where we do apply Assumption 9(a), we observe that the two solutions with Assumption 9(a) vary insignificantly for the MS, CG and DG reward structures.

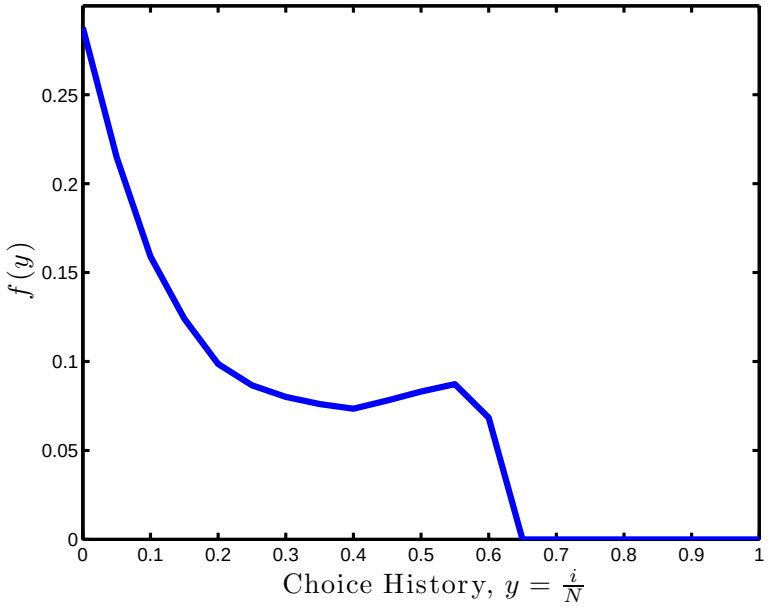


Figure 5.1: Average difference in anticipated reward $f(y) = w_B - w_A$ used in Assumption 9(b) for the RO reward structure (shown here for $N = 20$). When Assumption 9(b) applies, this finite set of values is substituted for Δr.

Assumption 9(b) is specific to the RO reward structure of Figure 3.1(b) where Assumption 9(a) does not apply. The function $f(y)$ is determined from a simulation in which the model decision maker made choice sequences in the RO task and the computed anticipated rewards for each value of y were averaged. In the simulation $\lambda = 0.1$, which approximates well the fitted value of λ for this RO structure [57]. Figure 5.1 is a plot of $f(y)$ as a function of $y = \frac{i}{N}, i = 1, 2, \ldots N$ in the case that $N = 20$. Note that for $y \geq 0.65$, $f(y) = 0$. A decision maker faced with the RO reward structure in the alone condition does not make choice A often enough to achieve $y(t) \geq 0.65$. This is directly related to the difficulty of the RO task.

5.2 Markov Model

In this Section we define the Markov process which is used to study the soft-max choice model in the TAFC task. Proposition 4 defines transition probabilities for a process with state $y(t)$ that evolves according to the strategy defined by the soft-max model.

Proposition 4. *Suppose Assumptions 8 and 9(a) hold. Then, the model decision maker (3.7) for the TAFC task (3.1)-(3.3) is a Markov process with state $y(t)$ and transition probabilities given by*

$$\Pr\{y(t+1) = y(t) - \frac{1}{N}\} = \frac{e^{\mu\Delta r} y(t)}{1 + e^{\mu\Delta r}} \tag{5.1}$$

$$\Pr\{y(t+1) = y(t)\} = \frac{e^{\mu\Delta r} + (1 - e^{\mu\Delta r})y(t)}{1 + e^{\mu\Delta r}} \tag{5.2}$$

$$\Pr\{y(t+1) = y(t) + \frac{1}{N}\} = \frac{1 - y(t)}{1 + e^{\mu\Delta r}} \tag{5.3}$$

where $\Delta r = \Delta r(y(t))$ is given by (3.5). In case Assumption 9(b) holds instead of Assumption 9(a), then the transition probabilities are given by (5.1)-(5.3) with $\Delta r(y(t))$ replaced with $f(y(t))$.

Proof of Proposition 4:

Since for a given choice $x_1(t+1)$ at time $t+1$, $y(t+1)$ can only change from its current value of $y(t)$ to $y(t) + \frac{1}{N}$, $y(t) - \frac{1}{N}$ or stay at $y(t)$, we need only compute the probability of each of these three events for all $y(t) \in \mathcal{Y}$. Each of these events depends upon the current value of $y(t)$ as well as $x_1(t+1)$ and $x_N(t)$ since $y(t+1)$ will only differ from $y(t)$ if $x_1(t+1)$ also differs from $x_N(t)$.

The event that $y(t+1) = y(t) - \frac{1}{N}$ requires $x_1(t+1) = B$ and $x_N(t) = A$. Treating these as independent events and using (3.7) with Assumption 8 yields

$$\begin{aligned} \Pr\{y(t+1) = y(t) - \frac{1}{N}\} &= \Pr\{x_1(t+1) = B\}\Pr\{x_N(t) = A\} \\ &= \frac{e^{\mu(w_B(t) - w_A(t))} y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}. \end{aligned}$$

Substituting in the identity of Assumption 9(a), we get (5.1).

Similarly, the probability that $y(t+1)$ takes the value $y(t) + \frac{1}{N}$ is given by

$$\begin{aligned} \Pr\{y(t+1) = y(t) + \frac{1}{N}\} &= \Pr\{x_1(t+1) = A\}\Pr\{x_N(t) = B\} \\ &= \frac{1 - y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}. \end{aligned}$$

Substituting in the identity of Assumption 9(a), we get (5.3).

The event that $y(t+1) = y(t)$ requires either $x_1(t+1) = A$ and $x_N(t) = A$ or $x_1(t+1) = B$ and $x_N(t) = B$. The probability of the union of these events is

$$\begin{aligned} \Pr\{y(t+1) = y(t)\} &= \Pr\{x_1(t+1) = A\}\Pr\{x_N(t) = A\} \\ &+ \Pr\{x_1(t+1) = B\}\Pr\{x_N(t) = B\} \\ &= \frac{y(t) + (1 - y(t))e^{\mu(w_B(t) - w_A(t))}}{1 + e^{\mu(w_B(t) - w_A(t))}}. \end{aligned}$$

Substituting in the identity of Assumption 9(a), we get (5.2). Since all of the probabilities depend on $y(t)$ only, the state at time t, the process is Markov. The case when Assumption 9(b) holds follows similarly. In that case, the finite set of values (shown in Figure 5.1) is substituted for Δr. □

Equations (5.1)-(5.3) are used to build the $(N+1) \times (N+1)$ state transition matrix $\mathbf{P}$ which has entries $P_{ij} = \Pr\{y(t+1) = \frac{j}{N} | y(t) = \frac{i}{N}\}, i, j \in \{0, 1, \ldots, N+1\}$.

5.3 Steady-State Choice Distribution

By deriving a steady-state choice distribution for the Markov process we have identified, we develop an analytic expression for the fraction of time a decision maker spends at each state in the task. The steady-state choice distribution serves as a predictive tool. Using the expression derived in this section, we derive performance as a function of the relevant parameters, study sensitivity of the model to those parameters, and even make predictions about decision making in scenarios that have not yet been studied.

Since the Markov process modeled in Section 5.2 is tridiagonal with strictly positive elements, any state can be reached from any another in finite time, guaranteeing irreducibility. It is aperiodic since return to state i from state i can happen as quickly as one time step, but no state is absorbing. Thus, the process has a unique limiting distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_N)$ describing the fraction of time the chain will spend in each of the enumerated states $y = i/N$, $i = 0, 1, \ldots, N$, in the long run (as $t \to \infty$) [83]. This steady-state distribution is the solution to the following equations:

$$\pi \mathbf{P} = \pi \tag{5.4}$$

$$\sum_{i=0}^{N} \pi_i = 1. \tag{5.5}$$

We derive this steady-state distribution with the following proposition.

Proposition 5. *For the transition probabilities given by (5.1) - (5.3) the unique steady-state distribution is*

$$\pi_i = \frac{\alpha_i (1 + e^{\mu \Delta r (\frac{i}{N})}) e^{-\mu \beta_i}}{\sum_{j=0}^{N} \alpha_j e^{-\mu \beta_j} (1 + e^{\mu \Delta r (\frac{j}{N})})} \tag{5.6}$$

where $\alpha_i = \frac{N!}{(N-i)!i!}$ *and* $\beta_i = \sum_{j=1}^{i} \Delta r(\frac{j}{N})$.

Proof of Proposition 5: Solving (5.4) alone yields a row vector v with elements given by

$$v_i = \frac{N!}{(N-i)!i!}(1 + e^{\mu \Delta r(\frac{i}{N})})e^{-\mu \sum_{j=1}^{i} \Delta r(\frac{j}{N})}.$$

To solve (5.5) we normalize the vector v to get

$$\pi = \frac{v}{\sum_{i=0}^{N} v_i}.$$

The elements of π are then given by (5.6). □

The distribution π from (5.6) is plotted in Figure 5.2 for each of the reward structures shown in Figure 3.1. The distribution in Figure 5.2(a) shows that the decision maker in the MS task spends most of the time making choices that keep y near the matching point (rather than near the optimal solution at the peak of the average reward curve). Should choice A be made more frequently, then higher rewards would be received.

Figure 5.2(b) shows the distribution for a decision maker in the RO task. This plot illustrates that, when faced with the RO reward structure, the decision maker is unable to find the global optimum at $y = 1$. Instead, time is spent at the local optimum where $y = 0$ and near the matching point. It is indeed rare for a decision maker to find the global optimum in this RO task [57].

In contrast Figure 5.2(c) shows for the CG structure that the decision maker spends most time making choices that yield the optimal average reward. Note that the distribution peaks at the matching point; i.e. the decision maker spends the highest fraction of time with proportion of choice A at the matching point. This is a much easier task since the optimum coincides with the matching point. The DG task is more difficult than the CG task since the optimum is divergent. The symmetry

of the reward structures causes decision makers to diverge to choice sequences with $y < 0.5$ and $y > 0.5$ with equal probability. The distribution in Figure 5.2(d) shows how choice sequences diverge from the point of intersection in the reward curves. The optimal strategy is for a decision maker to choose A half the time - as is typical of decision makers in the CG task.

The distribution shown in Figure 5.2(a) agrees qualitatively with experimental results, see for example those reported in [28]. Likewise the predicted distributions in Figures 5.2(b), 5.2(c) and 5.2(d) agree qualitatively with experimental results in [56, 57]. We also include direct comparisons to experimental data in Section 5.5.

5.4 Performance

The analytic expression π for the steady-state behavior presented in 5.6 leads to a number of interesting findings. In this section we look at performance of the soft-max model in the TAFC task through two approaches.

First we show conditions for which the decision maker converges to a value of y that corresponds to a matching point. To do this we consider a class of reward structures with a unique matching point (MS and CG reward structures of Figures 3.1(a) and 3.1(c) are two examples). Second we derive the sensitivity of the expected value of reward earned by the decision maker to the parameter μ in the soft-max model (3.7). This can be done by direct differentiation of (5.6). The sensitivity analysis appears in Section 5.4.2.

5.4.1 Steady-State Matching

For clarity, we again define the concept of reward structures with a matching point.

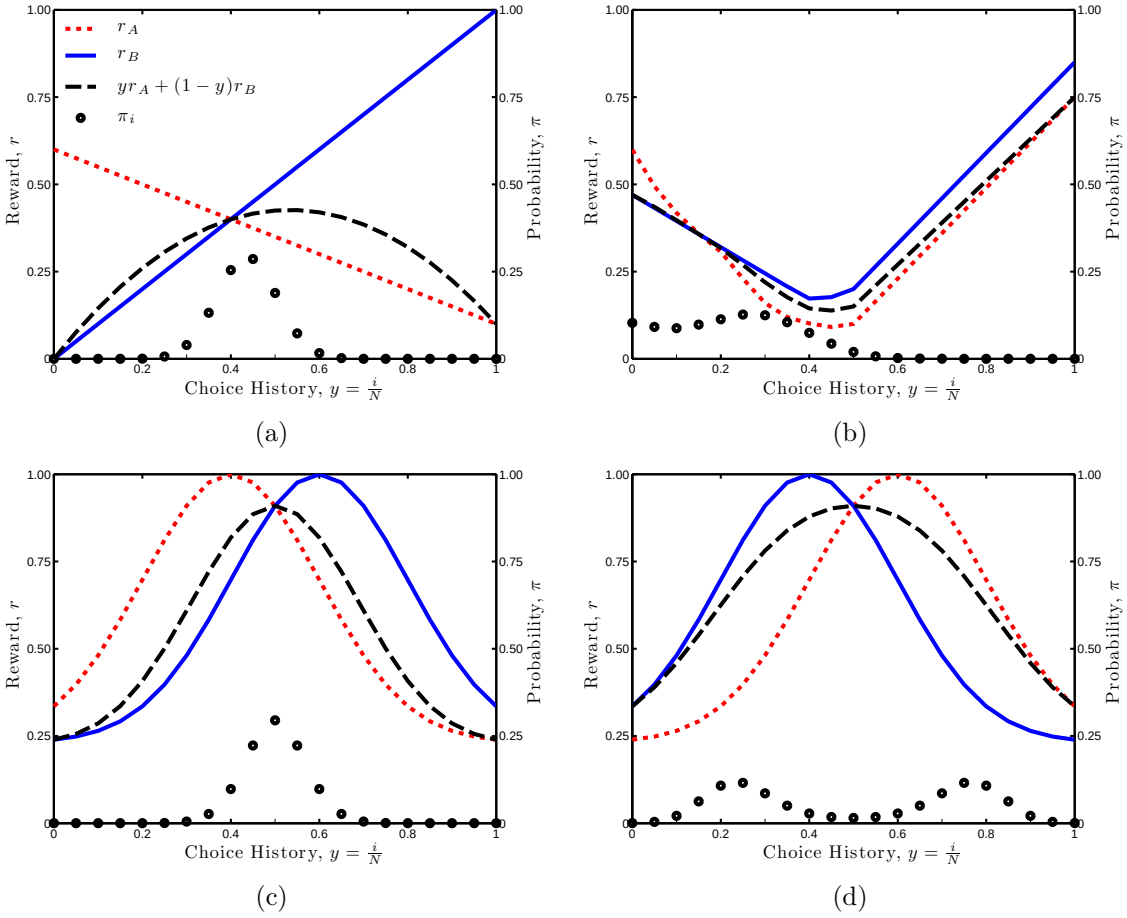


Figure 5.2: Steady-state distributions: The probability π_i that the decision maker has proportion $y = \frac{i}{N}$ of A's in their choice history, as given by (5.6), is shown for $i = 1, 2, \ldots, N$ by the circular points for each reward structure with $N = 20$. As before, in each plot the dotted line depicts r_A, the reward for choice A. The solid line depicts r_B, the reward for choice B. The dashed line is the average value of the reward. Each is plotted as a function of y. The values of the parameter μ for RO, CG and DG tasks are from the best fit to experimental data [57]. (a) MS with $\mu = 5$, (b) RO with $\mu = 11$, (c) CG with $\mu = 2.5$, (d) DG with $\mu = 2.91$.

Definition 1. A reward structure with a unique matching point of type 1 *consists of reward curves* $r_A(y)$, $r_B(y)$ *for which there exists* $y^* = \frac{i^*}{N}$, $i^* \in \{1, 2, \ldots, N-1\}$, *that satisfy* $\Delta r(y^*) = 0$, $\Delta r(y) < 0$ *for* $y < y^*$, *and* $\Delta r(y) > 0$ *for* $y > y^*$.

There are relatively few results that prove conditions for matching behavior given well-established models like the soft-max choice model. This work represents the first rigorous mathematical proof of matching behavior for the soft-max choice model.

In this section we prove steady-state matching behavior for the soft-max model by finding sufficient conditions on the slope μ of the soft-max function that guarantee that π_i is greatest for $y = i/N$ at or near the matching point. In Theorem 4 below, we find a bound μ_1 such that if $\mu > \mu_1$ then π_i peaks in a small neighborhood of the matching point. In Theorem 5 we find a bound $\mu_2 > \mu_1$ such that if $\mu > \mu_2$ then π_i peaks at the matching point.

Theorem 4. *Consider a reward structure with a unique matching point of type 1 and suppose that Assumptions 8 and 9(a) hold. If*

$$\mu > \mu_1 := max\left\{ \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*+2}{N}\right)}, \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*-2}{N}\right)} \right\} \tag{5.7}$$

then the steady-state choice distribution is maximum for $y \in \{y^* - \frac{1}{N}, y^*, y^* + \frac{1}{N}\}$. *Here* $\gamma = \frac{(N-i^*)!i^*!}{2\lfloor \frac{N}{2} \rfloor! \lceil \frac{N}{2} \rceil!}$ *where* $\lfloor \cdot \rfloor$ *gives the largest integer less than its argument and* $\lceil \cdot \rceil$ *gives the smallest integer greater than or equal to its argument.*

Proof of Theorem 4: To prove Theorem 4 we examine $\rho(i) = \pi_i / \pi_{i^*}$, the ratio of time spent at $y = \frac{i}{N}, i \neq i^*$ to time spent at $y^* = \frac{i^*}{N}$. From (5.6) we compute

$$\rho(i) = \frac{(N-i^*)!i^*!(1 + e^{\mu \Delta r(\frac{i}{N})}) e^{-\mu \sum_{j=1}^{i} \Delta r(\frac{j}{N})}}{2(N-i^*)!i^*! e^{-\mu \sum_{j=1}^{i^*} \Delta r(\frac{j}{N})}}. \tag{5.8}$$

We show that $\rho(i) < 1$ for all $i \notin \{i^* - 1, i^*, i^* + 1\}$ by proving each of two cases. In the first case we show that $\rho(i) < 1$ for all $i > i^* + 1$. In the second case we show that $\rho(i) < 1$ for $i < i^* - 1$.

Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$. The ratio $\rho(i)$ then becomes

$$\rho(i) = \frac{(N - i^*)!i^*!(1 + e^{\mu\Delta r(\frac{i^*+\epsilon}{N})})e^{-\mu\left(\Delta r(\frac{i^*+1}{N})+\cdots+\Delta r(\frac{i^*+\epsilon}{N})\right)}}{2(N - i^* - \epsilon)!(i^* + \epsilon)!}. \tag{5.9}$$

Replacing $(N - i)!i!$ in the denominator of (5.9) with its minimal possible value for $i \in \{0, 1, \ldots, N\}$ yields the inequality

$$\rho(i) \leq \gamma(1 + e^{-\mu\Delta r(\frac{i^*+\epsilon}{N})})e^{-\mu\left(\Delta r(\frac{i^*+1}{N})+\cdots+\Delta r(\frac{i^*+\epsilon-1}{N})\right)} \tag{5.10}$$

where $\gamma = \frac{(N-i^*)!i^*!}{2\lfloor\frac{N}{2}\rfloor!\lceil\frac{N}{2}\rceil!}$.

Now assume $\epsilon \geq 2$. Since $\Delta r(\frac{i^*+\epsilon}{N}) > 0$ for all $\epsilon \geq 1$, $\rho(i)$ decreases with increasing ϵ so from (5.10) we can write the strict inequality

$$\rho(i) < \gamma(1 + e^{-\mu\Delta r(\frac{i^*+2}{N})}). \tag{5.11}$$

If (5.7) is satisfied then (5.11) becomes $\rho(i) < 1$.

Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. Following the same steps as in Case 1, and making use of the fact that $\Delta r(\frac{i^*-\epsilon}{N}) < 0$ for all $\epsilon > 0$, we can write the inequality

$$\rho(i) \leq \gamma(1 + e^{-\mu|\Delta r(\frac{i^*-\epsilon}{N})|})e^{-\mu\left(|\Delta r(\frac{i^*-\epsilon+1}{N})|+\cdots+|\Delta r(\frac{i^*}{N})|\right)}. \tag{5.12}$$

Now assume $\epsilon \leq -2$. Since $\rho(i)$ decreases with decreasing ϵ for $\epsilon < 0$, we can write the strict inequality

$$\rho(i) < \frac{(N-i^*)!i^*!}{\lfloor \frac{N}{2} \rfloor ! \lceil \frac{N}{2} \rceil !}(1 + e^{-\mu|\Delta r(\frac{i^*-2}{N})|}). \tag{5.13}$$

If (5.7) is satisfied then (5.13) becomes $\rho(i) < 1$. □

Theorem 5. *Consider a reward structure with a unique matching point of type 1 and suppose that Assumption 8 and 9(a) hold. If*

$$\mu > \mu_2 := max\left\{ \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*+1}{N}\right)}, \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*-1}{N}\right)} \right\} \tag{5.14}$$

then the steady-state choice distribution is maximum for $y = y^*$. Here $\gamma = \frac{(N-i^*)!i^*!}{2\lfloor \frac{N}{2} \rfloor ! \lceil \frac{N}{2} \rceil !}$ where $\lfloor \cdot \rfloor$ gives the largest integer less than its argument and $\lceil \cdot \rceil$ gives the smallest integer greater than or equal to its argument.

Proof of Theorem 5: Again we examine $\rho(i) = \pi_i / \pi_{i^*}$. To prove Theorem 5 we follow the same process used in Theorem 4. We show that $\rho(i) < 1$ for all $i \neq i^*$ by proving each of two cases. In the first case we show that $\rho(i) < 1$ for all $i > i^*$. In the second case we show that $\rho(i) < 1$ for $i < i^*$.

Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$.

We assume $\epsilon \geq 1$. We have shown that $\rho(i)$ decreases with increasing ϵ so using Equation 5.10 we can arrive at the strict inequality

$$\rho(i) < \frac{N-i^*}{2(i^*+1)}(1 + e^{-\mu \Delta r(\frac{i^*+1}{N})}). \tag{5.15}$$

If (5.14) is satisfied then (5.15) becomes $\rho(i) < 1$.

Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. We assume $\epsilon \leq -1$. Since $\rho(i)$ decreases with decreasing ϵ for $\epsilon < 0$, we can write the strict inequality

$$\rho(i) < \frac{(N-i^*)!i^*!}{\lfloor \frac{N}{2} \rfloor ! \lceil \frac{N}{2} \rceil !}(1 + e^{-\mu|\Delta r(\frac{i^*-1}{N})|}). \tag{5.16}$$

If (5.14) is satisfied then (5.16) becomes $\rho(i) < 1$. □

Example 1: For the MS reward structure of Figure 3.1(a), we have $r_A(y) = k_A y + c_A$ and $r_B(y) = k_B y + c_B$ where $k_A = -\frac{1}{2}, c_A = \frac{3}{5}, k_B = 1$ and $c_B = 0$. For this example with $N = 20$, by Theorems 4 and 5, $\mu_1 = 5.45$ and $\mu_2 = 10.91$. These values shrink for smaller N and grow for larger N.

Example 2: For the CG reward structure of Figure 3.1(c), we have

$$r_A(y) = e^{-\left(\frac{y-\bar{y}_A}{\sqrt{2}\sigma_A}\right)^2} + c_A, \qquad r_B(y) = e^{-\left(\frac{y-\bar{y}_B}{\sqrt{2}\sigma_B}\right)^2} + c_B \tag{5.17}$$

with $\bar{y}_A = \frac{2}{5}, \bar{y}_B = \frac{3}{5}$ and $\sigma_A = \sigma_B = \frac{1}{5}$ and $c_A = c_B = \frac{3}{10}$. In this example with $N = 20$, by Theorem 4, $\mu_1 = 3.30$. By Theorem 5, $\mu_2 = 6.06$.

5.4.2 Sensitivity to Model Parameters

A number of interesting quantities can be computed from the expression for π in Equation (5.6). The analyses in this chapter can be used to develop tools for design and performance metrics to monitor decision making, or make predictions before tasks are even performed. In the previous section we proved conditions for matching. Here we are concerned with performance in the long run.

Given the fraction of time spent at each proportion of choice A, we can compute sensitivity of long-run performance to task parameters. This sensitivity is computed

here with respect to the parameter μ in the soft-max function. As mentioned in Section 3.4.2, larger μ corresponds to increased certainty in the decision making, which can also be interpreted as a reduced tendency to explore.

The average reward, from Equation (3.4), can be written $\bar{r}(y) = yr_A(y) + (1 - y)r_B(y)$. For each value of y, this is the reward that would be received on average if the decision maker were to maintain that value of y. So the expected value of the reward is the sum of each average reward multiplied by the fraction of time spent at each corresponding proportion of choice A and is written

$$\tilde{r} = \sum_{i=0}^{N} \pi_i \bar{r}_i. \tag{5.18}$$

The sensitivity of performance to μ can then be computed as the derivative of the expected value of the reward with respect to μ:

$$\frac{d}{d\mu}\tilde{r} = \sum_{i=0}^{N} \bar{r}_i \frac{d}{d\mu}\pi_i = \sum_{i=0}^{N} \left(\frac{i}{N} r_A\big(\frac{i}{N}\big) + \frac{N-i}{N} r_B\big(\frac{i}{N}\big) \right) \frac{d}{d\mu}\pi_i. \tag{5.19}$$

By denoting $g_i(\mu) := (1 + e^{\mu \Delta r(\frac{i}{N})})$ and $M(\mu) := \sum_{j=0}^{N} \pi_j$, the derivative of π_i with respect to μ can be written

$$\begin{aligned} \frac{d}{d\mu}\pi_i = & \frac{\alpha_i e^{-\mu\beta_i}(\Delta r(\frac{i}{N}) e^{\mu \Delta r(\frac{i}{N})} - \beta_i g_i(\mu))}{M(\mu)} \\ & - \frac{\alpha_i e^{-\mu\beta_i} g_i(\mu) \sum_{j=0}^{N} \alpha_j e^{-\mu\beta_j} \big(\Delta r(\frac{j}{N}) e^{\mu \Delta r(\frac{j}{N})} - \beta_j g_j(\mu)\big)}{M(\mu)^2}. \end{aligned} \tag{5.20}$$

Example 1 continued: Consider again the MS reward structure of Figure 3.1(a). The derivative of the expected value of reward with respect to μ, given by (5.20) is plotted in Figure 5.3 along with the expected value of the reward for $N = 20$. For this reward structure there is a critical point at $\mu_c = 1.15$. For $\mu < \mu_c$ increasing μ

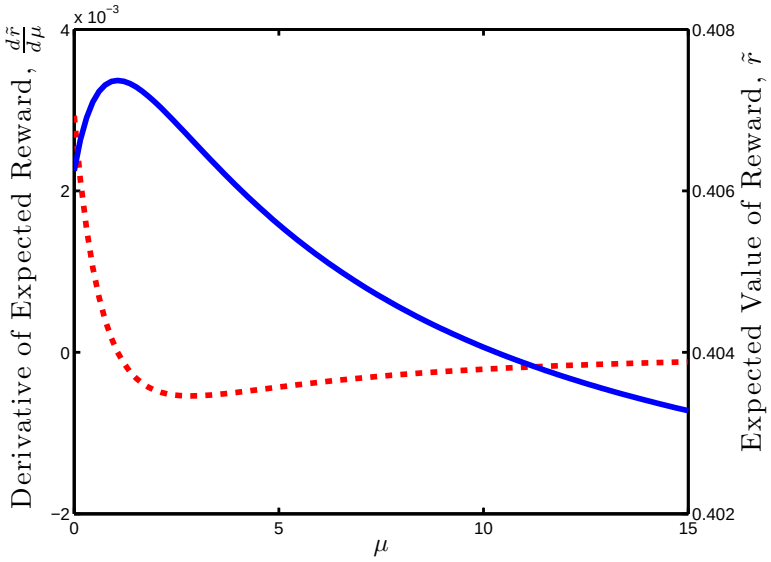


Figure 5.3: The derivative of the expected value of reward for the MS reward structure shown in Figure 3.1(a). The dotted line is $\frac{d}{d\mu}\tilde{r}$ from Equation (5.20). The solid line is the expected value of reward, $\tilde{r}$. Both are plotted as a function of μ for $N = 20$.

results in substantially higher reward. However, as μ increases further, the expected value of reward decreases. This is an example for which some exploratory behavior in the decision making is beneficial and is directly related to the results of Theorems 4 and 5: for $\mu > \mu_1 = 5.11$, i.e., with too much certainty (equivalently not enough exploration), the decision maker converges to the matching point, which is not the optimal strategy.

Example 2 continued: Consider again the CG reward structure of Figure 3.1(c). The derivative of the expected value of reward with respect to μ, given by (5.20) is plotted in Figure 5.4 along with the expected value of the reward for $N = 20$. In this example, $\frac{d}{d\mu}\tilde{r}$ is positive for all μ. The derivative is always positive in this example (for any N) because the matching point coincides with the maximum of the expected value of reward.

Whenever a decision maker converges to y^* in the CG reward structure, it is also the case that the highest reward on average is received. Therefore, increasing the parameter μ, or the certainty in the decision making, results in higher expected

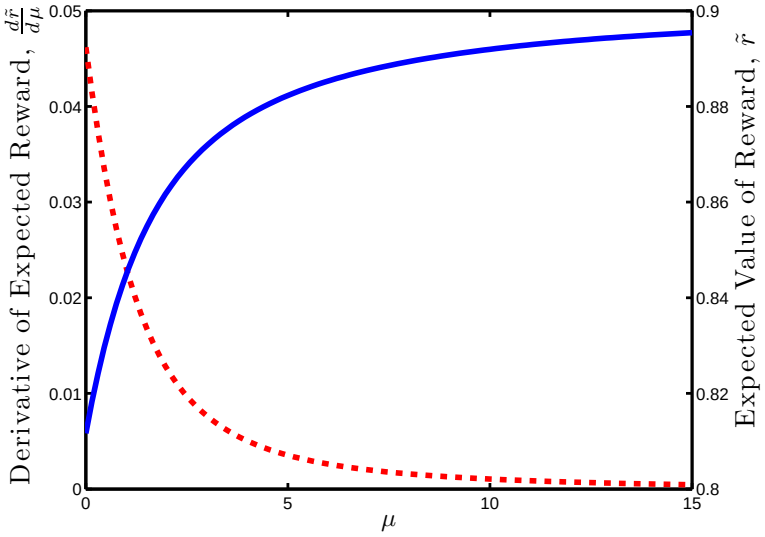


Figure 5.4: The derivative of the expected value of reward for the CG reward structure shown in Figure 3.1(c). The dotted line is $\frac{d}{d\mu}\tilde{r}$ from Equation (5.20). The solid line is the expected value of reward, $\tilde{r}$. Both are plotted against μ for $N = 20$.

reward for the task. We note, however, that there is not a great deal of gain in performance once μ increases above a threshold approximately equal to 5.

The expressions derived in this chapter can now serve as a tool for designing automated decision makers that use the soft-max choice model as a strategy. Designing automated decision makers becomes important in Chapters 6 and 7 where we consider mixed teams of decision makers in a network, each receiving social feedback.

5.5 Comparison to experimental results

To validate the analysis of our model, and to understand its limitations, we compare the steady-state distributions given by (5.6) to distributions taken from experimental data used in [56]. In these experiments subjects made a total of 150 choices. The distributions, plotted for the RO, CG, and DG reward structures in Figure 5.5, are the percentage of time spent with each possible choice history $y = \frac{i}{N}, i = 1, 2, \ldots, N$, averaged over all subjects in the alone (no social feedback) condition. In these experiments $N = 20$. Note that data for the MS structure is not shown here. The authors

of [56] and [57] do not run experiments for the MS task. For experimental data that shows convergence to the matching point in MS reward curves, and agrees well with our results, see Figure 2 of [28].

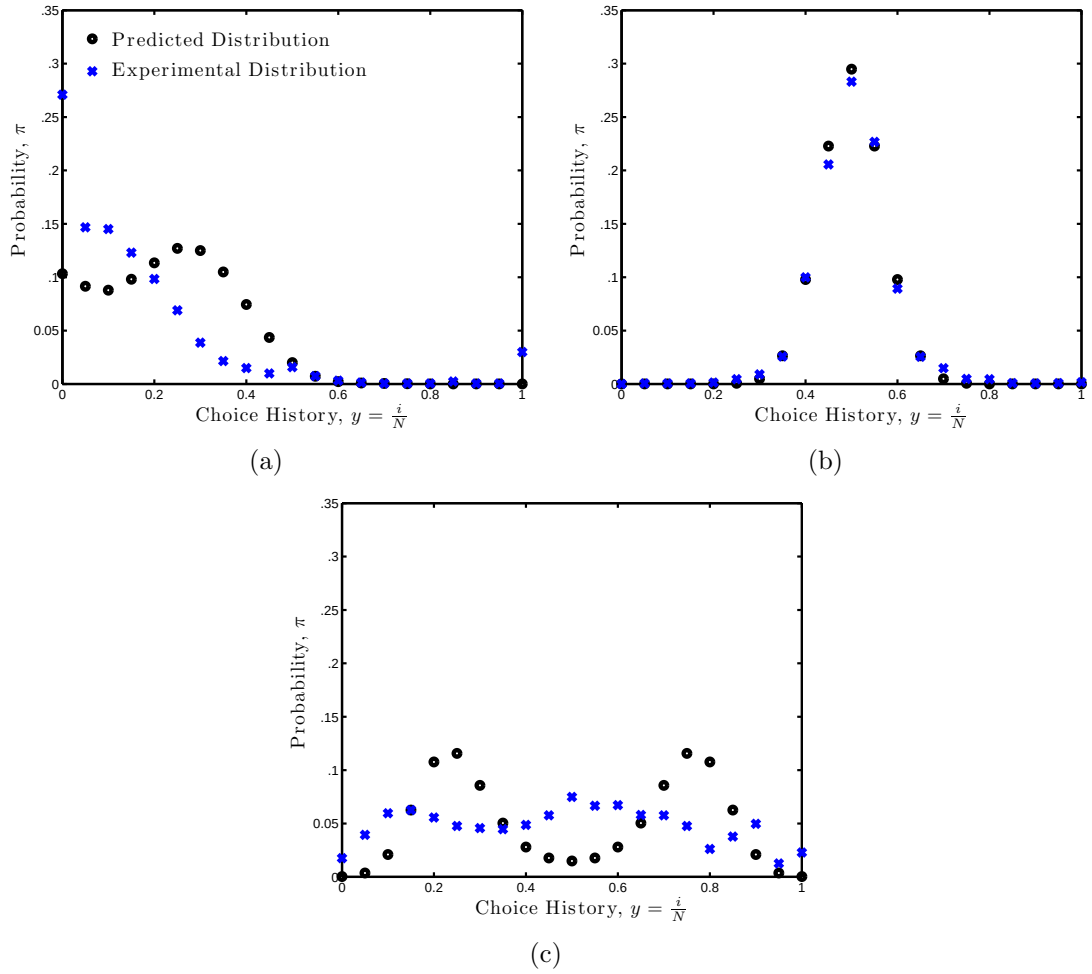


Figure 5.5: Comparison of experimental data from [56] to analytic prediction of the steady-state distrubution (5.6). Subjects made a total of 150 choices in the experiment and the distributions shown are averaged over all subjects. In each plot circles are from Equation (5.6) and x's are from data. (a) RO, (b) CG, (c) DG.

Some of the distributions shown in Figure 5.5 agree well with experimental data through a qualitative comparison. Some distributions agree with the data better than others. The distributions for the CG task in Figure 5.5(b) are very close to one another. The predicted distribution only differs slightly from the one computed from data. The distributions for the DG task in Figure 5.5(c), however, are dissimilar

The disagreement between predicted and computed distributions is likely due to the convergence rate which we understand varies among tasks. With only 150 choices made in these experiments, comparison between a steady-state distribution (derived for $t = \infty$), and experimental data depends upon the rate that subjects converge in each task. For instance, in the DG task, sufficient time was not allowed for subjects to converge to the steady-state distribution. In the CG task, however, the convergence rate is much higher. This is related to the difficulty level of each task (the CG task is easy and decision makers quickly find the optimum). Subjects typically do a lot of exploring in the more difficult DG task, in part due to the divergent nature of the optimal point in the reward structure.

We also see that the distributions for the RO task in Figure 5.5(a) differ in two ways. The distribution given by (5.6) is qualitatively similar to the distribution from data, but shifted slightly to the right. It is also true that subjects spent more time with $y = 0$ than predicted for the steady-state distribution. It's most important, however, to note that our analysis does predict that few subjects will discover the global optimum in the RO task; both distributions are close to zero for $y > 0.5$. It is likely that differences between (5.6) and experimental data for the RO task arise due to limitations on our ability to estimate $f(y)$ in Assumption 9(b). Since Assumption 9(a) does not apply, and we cannot directly measure $w_B - w_A$ for a subject, we find it convenient to estimate $f(y)$ using a maximum likelihood method. It is likely that this introduces some error which accounts for the differences in the curves plotted in Figure 5.5(a).

Differences in the distributions of Figure 5.5 emphasize the importance of properly applying our predictive tools. Perfect prediction of human behavior is not feasible, but understanding characteristics of the behavior and the role of relevant parameters is highly valuable. We aim to use these tools to assist in the design of mixed teams of decision makers. For example, in Chapter 6 we predict whether or not decision

makers in the RO task can be led toward the global optimum by introducing designed elements. Whether predictions of this type are suitable is an issue addressed in Chapter 7 where we discuss experiments that will test our hypotheses.

This concludes the analysis of this chapter. In the following chapter a similar approach is used to analyze decision making with social feedback. The method used in this chapter is extended to consider coupled model decision makers that share choice feedback.

Chapter 6

Decision Making in a Social Context and with Mixed Teams

In this chapter we leverage experimental results and extend the analysis of Chapter 5 to study coupling of soft-max model decision makers. We consider a group of model decision makers simultaneously making decisions with choice feedback in the TAFC task and develop a formal understanding of the role of feedback, network topology, and individual decision-making characteristics in the performance of individuals in the group. We derive, analytically, the steady-state distribution analogous to the expression for π that we develop in Chapter 5. We do so by considering the TAFC task in a social context as described in Section 3.3, and using the extended model to consider choice feedback as described in Section 3.4.4. The analysis is similar to that of Chapter 5 in that we employ the assumptions of Section 5.1 and again identify a Markov process around the state of a focal decision maker.

For the purpose of the tractability of our analysis, we focus on directed information passing, and we examine the decision dynamics of the focal individual who receives choice feedback from M others. Those M others do not receive any social feedback.

We also perform an analysis that includes bi-directional choice feedback, computing the steady-state distribution in that case numerically.

In Section 6.1 we identify the task and decision making-model as an inhomogeneous Markov process and we compute the expected state transition probabilities. We investigate convergence in Section 6.2 and compute the expected equilibrium probability distribution for the process. In Section 6.3 we study performance of decision making in the CG task in this social context. We first prove that our model predicts the (small) negative impact of choice feedback on performance in the CG task as observed in [57]. We next study the sensitivity of performance in the CG task to decision-making parameters μ, ν and M. In Section 6.4 we analyze decision-making dynamics in the case of undirected social feedback and compare results for undirected versus directed feedback in the CG and DG tasks. In Section 7.1, we make predictions on the role of choice feedback in the difficult RO task by designing a heterogeneous group of M others and investigating conditions that lead to dramatic improvement in performance of the focal decision maker.

The approach taken in this chapter to use a Markov chain for analyzing the softmax choice model is analogous to that of Chapter 5. This method has been published with preliminary results in [78] and [80]. We focus primarily on social effects in [80]. The work that appeared in [80] was presented at the IEEE Conference on Decision and Control in Atalanta, GA in 2010. The results presented in this chapter have been published in [79].

6.1 Expectation of the Markov Model

The identification of the dynamics of the focal decision maker with directed choice feedback from M others as a Markov process with state $y(t)$ is analogous to that for the decision maker in the alone condition studied in Section 5.2. However, in the

social context the Markov process is inhomogeneous because at each time t, the one-step state transition matrix depends upon the choices of others through the function $u(t)$ in (3.21), which models the social feedback imposed by the choices condition. We can calculate the probability of each possible outcome for the value of $u(t)$ at each time. By conditioning on the value of $u(t)$ we analyze the expectation of the inhomogeneous process. The expectation of the state transition matrix is given in the following proposition.

Proposition 6. *Suppose Assumptions 8 and 9(a) hold. Then, the model representing the focal individual receiving choice feedback from M others (3.20)-(3.21) for the TAFC task (3.1)-(3.3) is a Markov process with state $y(t)$ and expected state transition probabilities given by*

$$\Pr\{y(t+1) = y(t) - \frac{1}{N}\} = \big[1 - \bar{p}_A\big(y(t)\big)\big]y(t) \tag{6.1}$$

$$\Pr\{y(t+1) = y(t)\} = \big[1 - \bar{p}_A\big(y(t)\big)\big]\big(1 - y(t) + \bar{p}_A\big(y(t)\big)y(t) \tag{6.2}$$

$$\Pr\{y(t+1) = y(t) + \frac{1}{N}\} = \bar{p}_A\big(y(t)\big)\big(1 - y(t)\big) \tag{6.3}$$

where $\Delta r = \Delta r(y(t))$ is given by (3.5) and $\bar{p}_A(y(t), \nu)$ is

$$\bar{p}_A(y(t), \nu) = \frac{Pr\{u(t) = 1\}}{1 + e^{\mu(\Delta r - \nu)}} + \frac{Pr\{u(t) = -1\}}{1 + e^{\mu(\Delta r + \nu)}} + \frac{Pr\{u(t) = 0\}}{1 + e^{\mu \Delta r}}. \tag{6.4}$$

The conditional probabilities on $u(t)$ are given by

$$\Pr\{u(t)=1\} = \sum_{k=\lceil \frac{M+1}{2} \rceil}^{M} \left(\begin{array}{c} M \\ k \end{array} \right) p_A(\infty,0)^k (1-p_A(\infty,0))^{M-k},$$

$$\Pr\{u(t)-1\} = \sum_{k=\lceil \frac{M+1}{2} \rceil}^{M} \left(\begin{array}{c} M \\ k \end{array} \right) (1-p_A(\infty,0))^k p_A(\infty,0)^{M-k},$$

$$\Pr\{u(t)=0\} = 1 - \big(Pr\{u(t)=1\} + Pr\{u(t)=-1\}\big) \tag{6.5}$$

and $\left(\begin{array}{c} M \\ k \end{array} \right) = \frac{M!}{k!(M-k)!}$. *In case Assumption 9(b) holds instead of Assumption 9(a), then the results hold with* $\Delta r(y(t))$ *replaced with* $f(y(t))$.

Proof of Proposition 6:

Since for a given choice $x_1(t+1)$ at time $t+1$, $y(t+1)$ can only change from its current value of $y(t)$ to $y(t)+\frac{1}{N}$, $y(t)-\frac{1}{N}$ or stay at $y(t)$, we need only compute the probability of each of these three events for all $y(t) \in \mathcal{Y}$. Each of these events depends upon the current value of $y(t)$ as well as $x_1(t+1)$ and $x_N(t)$ since $y(t+1)$ will only differ from $y(t)$ if $x_1(t+1)$ also differs from $x_N(t)$.

The event that $y(t+1) = y(t) - \frac{1}{N}$ requires $x_1(t+1) = B$ and $x_N(t) = A$. Treating these as independent events and using (3.20) yields

$$\Pr\{y(t+1) = y(t) - \frac{1}{N}\} = \Pr\{x_1(t+1)=B\} * \Pr\{x_N(t)=A\} = \frac{e^{\mu(w_B(t)-w_A(t)-\nu u(t))}y(t)}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.$$

Substituting in the identity of Assumption 9(a), and treating the M peer decisions as independent events, we condition on the value of $u(t)$ and get $\Pr\{x_1(t+1)=B\} = 1-\bar{p}_A(y(t),\nu)$ which with Assumption 8 gives us (6.1).

Similarly, the probability that $y(t+1)$ takes the value $y(t)+\frac{1}{N}$ is given by

$$\Pr\{y(t+1)=y(t)+\frac{1}{N}\}=\Pr\{x_1(t+1)=A\}*\Pr\{x_N(t)=B\}=\frac{1-y(t)}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.$$

Conditioning on the value of $u(t)$ and substituting in the identity of Assumption 9(a), we get (6.3).

The event that $y(t+1)=y(t)$ requires either $x_1(t+1)=A$ and $x_N(t)=A$ or $x_1(t+1)=B$ and $x_N(t)=B$. The probability of the union of these events is

$$\begin{aligned}\Pr\{y(t+1)=y(t)\}&=(\Pr\{x_1(t+1)=A\})(\Pr\{x_N(t)=A\})\\&\quad+\Pr\{x_1(t+1)=B\}*\Pr\{x_N(t)=B\}\\&=\frac{y(t)+(1-y(t))e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.\end{aligned}$$

Conditioning on the value of $u(t)$ and substituting in the identity of Assumption 9(a), we get (6.2). Since the probabilities depend only upon $y(t)$, the current value of the state at time t, the process is Markov. By conditioning on $u(t)$, the results (6.1)-(6.3) provide the expectation of the transition probabilities. The case when Assumption 9(b) holds follows similarly. □

The $(N+1)\times(N+1)$ one-step state transition matrix $P(t)$ has entries

$$P_{ij}=\Pr\{y(t+1)=\frac{j}{N}|y(t)=\frac{i}{N}\},\tag{6.6}$$

$i,j\in\{0,1,\ldots,N+1\}$. Using (6.1)-(6.3) we can build the expectation $\mathbf{P}$ of this transition matrix.

6.2 Convergence and Steady-State Choice Distribution

By identifying a Markov chain in this way for coupled soft-max model decision makers with choice feedback allows us to compute the steady-state distribution of y. In this section, we compute the expected steady-state distribution of y by using the expected state transition matrix $\mathbf{P}$. The process is similar to that of Chapter 5 where we considered isolated decision makers that did not receive social feedback.

Since the Markov process in Section 6.1 has a tridiagonal one-step state transition matrix with strictly positive elements, any state can be reached from any another in finite time, guaranteeing irreducibility. It is aperiodic since return to state i from state i can happen as quickly as one time step, but no state is absorbing. Thus, the process has a unique limiting distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_N)$ describing the fraction of time the chain will spend in each of the enumerated states $y = \frac{i}{N}, i = 0, 1, 2 \ldots N$, in the long run (as $t \to \infty$) [83]. We summarize this in the following proposition.

Proposition 7. *For the expected transition probabilities given by (6.1) - (6.3) the unique expected steady-state distribution is*

$$\pi_i = \alpha_i \frac{\prod_{j=1}^{i} q(\frac{i}{N}, \nu)}{\sum_{j=0}^{N} \alpha_j \prod_{k=1}^{j} q(\frac{k}{N}, \nu)} \tag{6.7}$$

where $\alpha_i = \frac{N!}{(N-i)!i!}$ *and* $q(\frac{i}{N}, \nu) = \frac{\bar{p}_A(\frac{i-1}{N}, \nu)}{1 - \bar{p}_A(\frac{i}{N}, \nu)}$.

Proof of Proposition 7: Solving (5.4) alone yields a row vector v whose elements are given by

$$v_i = \frac{N!}{(N-i)!i!} \prod_{j=1}^{i} \frac{\bar{p}_A(\frac{j-1}{N}, \nu)}{1 - \bar{p}_A(\frac{j}{N}, \nu)}.$$

To solve (5.5) we normalize the vector v to get $\pi = v/\sum_{i=0}^{N} v_i$. The elements of π are then given by (6.7). $\square$

6.3 Performance with Choice Feedback in the CG task

The CG reward structure is unique in that the maximum of the average reward (optimal point in the reward structure) coincides with an attracting matching point at $y^* = 0.5$. Therefore, decision makers in the CG task should maximize time spent with $y = 0.5$. For this reason, a key measure of performance in the TAFC task with CG reward structure is variance in y about $y^* = 0.5$. For the symmetric CG structures of the form (5.17) in which $c_A = c_B$, $\bar{y}_A < \bar{y}_B$ and $|\bar{y}_A| = |\bar{y}_B|$ (as in Figure 3.1(c)) better performance corresponds to minimizing the variance about $y = 0.5$.

In this section we prove for our predictive model of the expected steady-state distribution of $y(t)$ that the variance about $y = 0.5$ is minimal for $\nu = 0$; i.e., receiving choice feedback from other decision makers in the CG task is predicted to degrade performance. This agrees well with experimental results of [57]. We perform additional analyses in the section by investigating the effect on performance of the strength of the feedback ν, the number of other decision makers M and the focal individual's own exploratory parameter μ.

6.3.1 Effect of Choice Feedback on Reward

We prove here that variance is minimized when the strength of the choice feedback is smallest, i.e., when $\nu = 0$. The impact is that with choice feedback (corresponding to $\nu \neq 0$), the focal individual tends to do more exploring away from the optimal solution and performance deteriorates.

Let Σ denote the variance, or second moment, of the expected steady-state distribution about $y = 0.5$. Then using π given by (6.7), Σ can be written as a function of the feedback gain ν as

$$\Sigma(\nu) = \frac{\sum_{i=0}^{N} \alpha_i (\frac{i}{N} - \frac{1}{2})^2 Q_i(\nu)}{\sum_{j=0}^{N} \alpha_j Q_j(\nu)}, \tag{6.8}$$

where $Q_i(\nu) := \prod_{j=1}^{i} q(\frac{j}{N}, \nu)$.

The following theorem captures the result that choice feedback degrades performance in the TAFC task with CG reward structure.

Theorem 6. *Consider the CG reward structure of the form (5.17) where the matching point and optimal choice sequence coincide at $y = 0.5$. Consider a focal decision maker who receives choice feedback from $M = 4$ others who receive no feedback. Suppose that Assumptions 8 and 9(a) hold. Then, the variance $\Sigma(\mu, \nu)$ about $y = 0.5$ of the expected steady-state choice distribution of the focal individual is minimal for $\nu = 0$.*

We prove Theorem 6 by first proving four lemmas.

Lemma 16. *$\nu = 0$ is a critical point of $\Sigma(\mu, \nu)$*

To prove Lemma 16 we introduce the following:

Lemma 17. $Q_i'(0) := \frac{\partial}{\partial \nu} \prod_{j=1}^{i} q(\frac{i}{N}, \nu)\big|_{\nu=0} = 0.$

Proof of Lemma 17: We compute

$$\frac{\partial q}{\partial \nu}(\tfrac{i}{N}, \nu) = \frac{\frac{\partial}{\partial \nu}\bar{p}_A\big(\frac{i-1}{N}, \nu\big)}{1 - \bar{p}_A\big(\frac{i}{N}, \nu\big)} + \frac{\frac{\partial}{\partial \nu}\bar{p}_A\big(\frac{i}{N}, \nu\big)\big(1 - \bar{p}_A\big(\frac{i-1}{N}, \nu\big)\big)}{\big(1 - \bar{p}_A\big(\frac{i}{N}\big), \nu\big)^2}. \tag{6.9}$$

For the CG reward schedule $p_A(\infty, 0) = \frac{1}{2}$. Using this and $M = 4$ in (6.5), we can compute the conditional probabilities on $u(t)$. Substituting these into (6.4) for $\bar{p}_A$

gives

$$\bar{p}_A(\tfrac{i}{N},\nu) = \frac{3}{4}\frac{1}{(1+e^{\mu\Delta r})} + \frac{1}{8}\left[\frac{1}{1+e^{\mu(\Delta r-\nu)}} + \frac{1}{1+e^{\mu(\Delta r+\nu)}}\right] \tag{6.10}$$

where $\Delta r = \Delta r(\frac{i}{N})$. Differentiating $\bar{p}_A$ with respect to ν yields

$$\frac{\partial}{\partial\nu}\bar{p}_A(\tfrac{i}{N},\nu) = \frac{\mu e^{\mu\Delta r}}{8}\left[\frac{e^{-\mu\nu}}{(1+e^{\mu(\Delta r-\nu)})^2} - \frac{e^{\mu\nu}}{(1+e^{\mu(\Delta r+\nu)})^2}\right]. \tag{6.11}$$

Evaluating (6.11) at $\nu = 0$ we get $\frac{\partial}{\partial\nu}\bar{p}_A(\frac{i}{N},\nu)|_{\nu=0} = 0, \forall i$. Therefore, in (6.9) we see that $\frac{\partial}{\partial\nu}q(\frac{i}{N},\nu)|_{\nu=0} = 0$. From the definition of $Q_i(\nu)$ we can write

$$Q_i'(\nu) = \sum_{k=1}^{i}\frac{\partial}{\partial\nu}q\big(\tfrac{k}{N},\nu\big)\prod_{j=1,j\neq k}^{i} q\big(\tfrac{j}{N},\nu\big). \tag{6.12}$$

Evaluating (6.12) at $\nu = 0$ with $\frac{\partial}{\partial\nu}q(\frac{i}{N},\nu)|_{\nu=0} = 0$ gives $Q_i'(0) = 0$. □

Proof of Lemma 16: The derivative of $\Sigma(\mu,\nu)$ can be written

$$\frac{\partial}{\partial\nu}\Sigma(\mu,\nu) = \frac{\sum_{i=1}^{N}\alpha_i(\frac{i}{N}-\frac{1}{2})^2Q_i'(\nu)}{\sum_{k=1}^{N}\alpha_kQ_k(\nu)} - \frac{\sum_{i=1}^{N}\alpha_i(\frac{i}{N}-\frac{1}{2})^2Q_i(\nu)\sum_{k=1}^{N}\alpha_kQ_k'(\nu)}{\big(\sum_{k=1}^{N}\alpha_kQ_k(\nu)\big)^2}.$$

It follows from Lemma 17 that $\frac{\partial}{\partial\nu}\Sigma(\mu,\nu)|_{\nu=0} = 0$. □

It is now left to show that $\nu = 0$ is a minimum of $\Sigma(\mu,\nu)$.

Lemma 18. $\frac{\partial^2}{\partial\nu^2}\Sigma(\mu,\nu)\big|_{\nu=0} > 0$.

To prove Lemma 18 we introduce the following:

Lemma 19. $Q_i''(\nu) < 0$.

Proof of Lemma 19: Differentiating $Q_i'(\nu)$ with respect to ν, and making use of the fact that $\frac{\partial}{\partial\nu}\bar{p}_A(\frac{i}{N},\nu)|_{\nu=0} = 0$ gives

$$Q_i''(0) = \frac{\frac{\partial^2}{\partial\nu^2}\bar{p}_A(\frac{i-1}{N},\nu)|_{\nu=0}\big(1-\bar{p}_A(\frac{i}{N},0)\big)}{\big(1-\bar{p}_A(\frac{i}{N},0)\big)^2} + \frac{\frac{\partial^2}{\partial\nu^2}\bar{p}_A(\frac{i}{N},\nu)|_{\nu=0}\big(\bar{p}_A(\frac{i-1}{N},0)\big)}{\big(1-\bar{p}_A(\frac{i}{N},0)\big)^2}. \tag{6.13}$$

Since

$$\frac{\partial^2}{\partial \nu^2}\bar{p}_A(\frac{i}{N},\nu)|_{\nu=0} = -\frac{\mu^2 e^{\mu\Delta r(\frac{i}{N})}(1+e^{2\mu\Delta r(\frac{i}{N})})}{(1+e^{\mu\Delta r(\frac{i}{N})})^4} < 0, \tag{6.14}$$

we can conclude that $Q_i''(0) < 0$. $\square$

Proof of Lemma 18: Invoking Lemma 17 we can write

$$\frac{\partial^2}{\partial \nu^2}\Sigma(\mu,\nu)\big|_{\nu=0} = \frac{\sum_{i=0}^N \alpha_i\big(\frac{i}{N}-\frac{1}{2}\big)^2 Q_i''(0)}{\sum_{k=0}^N \alpha_k Q_k(0)} - \frac{\sum_{i=0}^N \alpha_i\big(\frac{i}{N}-\frac{1}{2}\big)^2 Q_i(0)\sum_{k=0}^N \alpha_k Q_k''(0)}{\big(\sum_{k=0}^N \alpha_k Q_k(0)\big)^2}. \tag{6.15}$$

Denote the numerator of $\frac{\partial^2}{\partial \nu^2}\big|_{\nu=0}\Sigma(\mu,\nu)$ by Γ. Then

$$\Gamma = \sum_{i=0}^N \sum_{k=0}^N \gamma_{i,k} \tag{6.16}$$

where $\gamma_{i,k} = \alpha_i Q_i''(0)\alpha_k Q_k(0)\big[\big(\frac{i}{N}-\frac{1}{2}\big)^2 - \big(\frac{k}{N}-\frac{1}{2}\big)^2\big]$. Lemma 19 tells us that $\gamma_{i,k} > 0$ for all i,k that satisfy

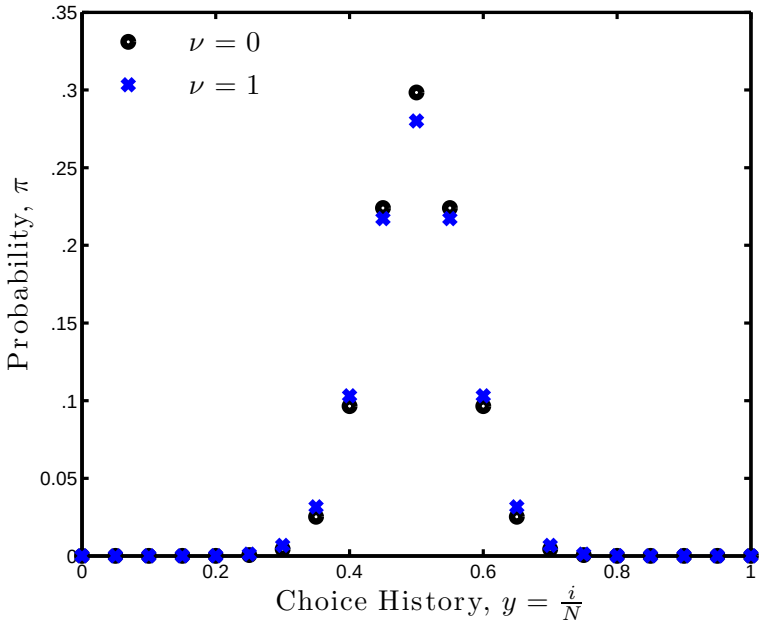
$$\left(\frac{i}{N}-\frac{1}{2}\right)^2 - \left(\frac{k}{N}-\frac{1}{2}\right)^2 < 0. \tag{6.17}$$

It is also true that $\gamma_{\frac{N}{2},\frac{N}{2}} = 0$. It can be shown that for all $i,k \neq \frac{N}{2}$ $\gamma_{i,k} > 0$, and that $\gamma_{\frac{N}{2},\frac{N}{2}} = 0$. It therefore must be true that $\Gamma = \sum_{i=0}^N \sum_{k=0}^N \gamma_{i,k} > 0$. $\square$

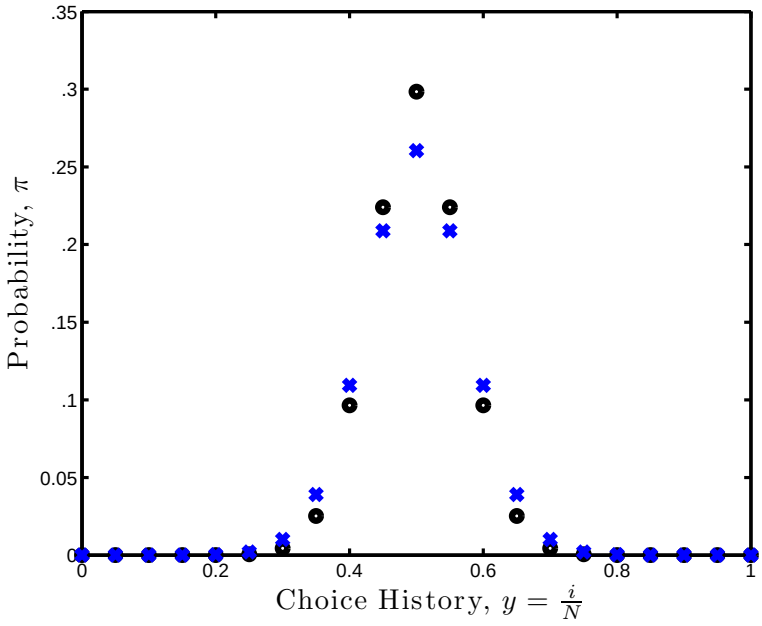
Proof of Theorem 6: Lemma 16 and Lemma 18 guarantee that $\nu = 0$ is a minimum of $\Sigma(\mu,\nu)$. $\square$

6.3.2 Sensitivity

With this model of choice feedback and corresponding steady-state distribution, we can compute sensitivities to parameters in the same fashion as was done in Chapter (5). The method involves differentiating Equation 3.20 with respect to the parameter of interest. In this section we examine the sensitivity of performance in the CG task with choice feedback to decision-making parameters ν, μ and M.



(a)



(b)

Figure 6.1: Steady-state distribution of y for the CG task with choice feedback. (a) $M = 4$, (b) $M = 2$. In each plot $\mu = 2.6$, the circles correspond to $\nu = 0$ (no feedback), and the x's to $\nu = 1$ ("full" feedback). The value of μ is chosen to be in accordance with the fitted values provided in [57].

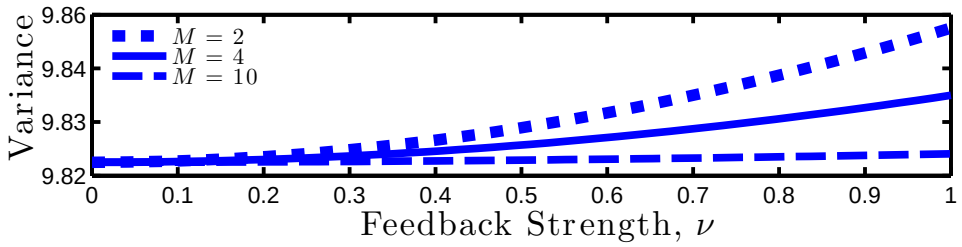
In Figure 6.1 the (expected) steady-state distribution of y is plotted without feedback and with feedback in the cases $M = 4$ and $M = 2$. Here we have used $\mu = 2.6$, which is the fitted value for an individual in the CG task with choice feedback [57]. In Figure 6.2 the normalized standard deviation $100\sqrt{\Sigma(\mu, \nu, M)}$ is plotted as a function of ν for three different values of μ and M. These results will be compared with experimental data in the following section.

In both figures, it can be seen that variance increases as a function of ν for each value of μ and M plotted; this is as predicted for the case $M = 4$ by Theorem 6. We also see that variance is higher for smaller M. This implies that social feedback of this kind has a greater effect on performance in smaller groups of decision makers.

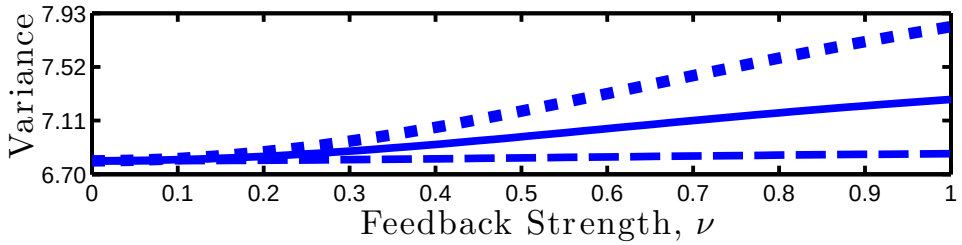
The results also show that dependence of the variance on μ is significant. In Figure 6.2 it can be seen that increasing μ (certainty in the decision making) magnifies sensitivity to the feedback gain ν. In Section 5.4.2 we showed that increasing μ in the CG task decreases variance for a single individual without social feedback. The exploratory parameter μ and the feedback gain ν have a coupling effect in the CG task with social feedback that causes a more substantial decrease in performance as ν increases for larger values of μ.

6.3.3 Comparison to Experimental Data

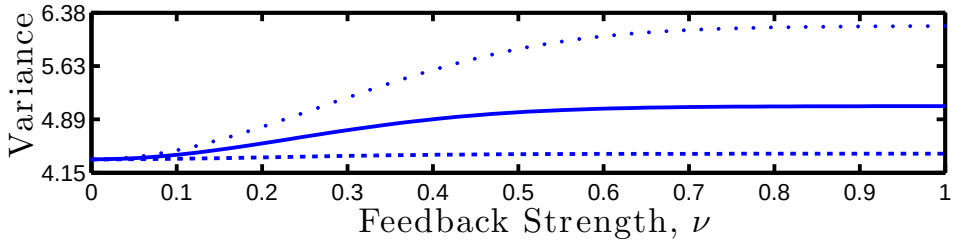
We have predicted that variance about the optimum in the CG task should increase with increasing influence of choice feedback. The authors of [56] and [57] have seen signs of this in the data, they also showed that performance deteriorates most significantly in the CG task when decision makers have social feedback that depends on the rewards of the M others, although we do not consider that feedback mode in this thesis. In Figure 6.3 we plot variance from the experiments performed in [57]. Here we are comparing variance from experimental data in two conditions: with choice feedback (dashed line) and without any feedback (solid line).



(a)



(b)



(c)

Figure 6.2: Standard deviation of expected steady-state distribution of y from the mean $y = 0.5$ for the CG task as a function of feedback parameter ν (given by $100\sqrt{\Sigma(\mu, \nu, M)}$). (a) $\mu = 0.5$, (b) $\mu = 2.60$, the fitted value from [57] (c) $\mu = 10$. In each plot, the dotted curve corresponds to $M = 2$, the solid curve to $M = 4$ and the dashed curve to $M = 10$.

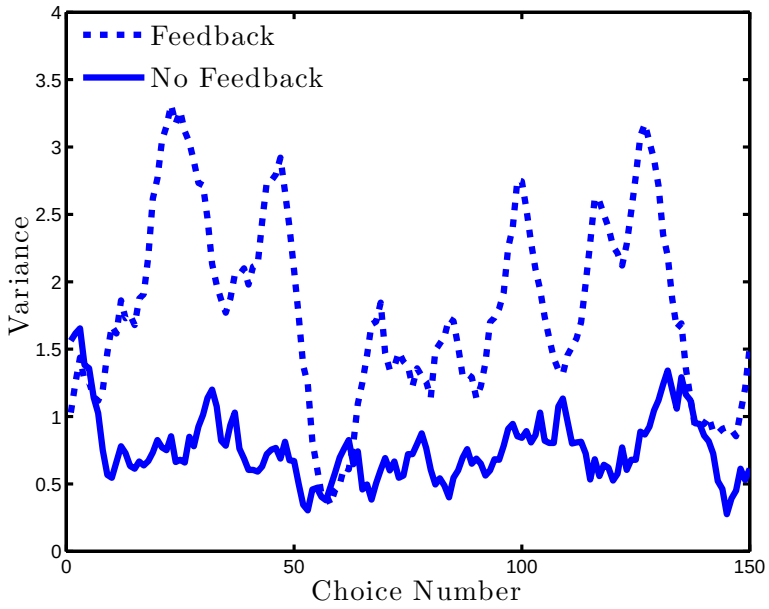


Figure 6.3: Variance in the decision making for experiments run with and without choice feedback. Each is plotted against the number of choices made. No feedback is shown with a solid line, choice feedback is shown with a dashed line. This figure was constructed using data provided by Andrea Nedic [57].

This data was collected from experiments with a duration of 150 choices and reward structures which depend on $N = 20$ of the most-recent choices. The variance is averaged over all subjects. There is a clear trend in Figure 6.3 which suggests that variance in the CG task with choice feedback is higher, thereby agreeing with our results in this section - in particular Theorem 6. It should be noted, however, that while the average variance over the 150 choices is higher with choice feedback, there are points in the data where variance in the alone condition is higher. It is possible that the effect is greater for longer time periods; i.e. approaching the steady state. It may also be the case that some subjects choose not to pay much attention to the social feedback, and therefore should be parametrized by a small ν.

6.4 Undirected Feedback

Conclusions drawn in Section 6.1-6.3 are the result of an analysis of decision making with directed social feedback. We are also interested in studying teams with social feedback that is undirected, particularly as undirected feedback was used in the experiments of [56, 57] (see Section 3.3). There may also be fundamental differences in the behavior (and performance) among teams with directed vs undirected feedback. A formal understanding of this difference can provide a valuable understanding for how to design the topology of networks used to build mixed teams.

It is not tractable to derive analytic results for coupled model decision makers that share choice feedback within a network that is not directed. The Markov chain required to analyze the undirected case prohibits us from deriving a closed-form, analytic solution for the steady-state distribution. In this section we numerically derive equilibrium distributions for a focal decision maker that receives, and shares, choice feedback in the TAFC task.

The method used here is to numerically compute steady-state distributions for a focal decision maker coupled with M others in an undirected network. We consider again a group of $(M+1)$ soft-max model decision makers simultaneously making decisions in the TAFC task in a social context. However, in the undirected case, each decision maker receives choice feedback from each of the other M decision makers, i.e., the graph that describes the communication topology is complete. The probability that any of the decision makers chooses A is given by (3.20) where feedback depends on the choices of others. We make Assumptions 8 and 9(a) (or 9(b)) so that the state of decision maker k is $y_k(t)$, $k = 1, \ldots, M+1$. Because the decision makers are all interconnected, we must retain the state of each decision maker, so the state of the system becomes $(y_1(t), \ldots, y_{M+1}(t))$.

To study the dynamics, we first identify the task and decision-making model as a Markov process. As in Section 6.1, the Markov process is inhomogeneous, and we

can compute the expectation of the state transition probabilities by conditioning on $u_k(t)$, $k = 1, \ldots, M+1$. We then use these probabilities to build the expected state transition matrix $\mathbf{P}$, which in this case will be a matrix of dimension $(N+1)^{M+1} \times (N+1)^{M+1}$.

Proposition 8. *Suppose Assumptions 8 and 9(a) hold. Then $(M+1)$ model decision makers each receiving choice feedback from the M others (3.20)-(3.21) for the TAFC task (3.1)-(3.3) form a Markov process with state $(y_1(t), \ldots, y_{M+1}(t))$ and expected state transition probabilities given by*

$$\Pr\{y_i(t+1) = y_i(t) + \frac{d_i}{N},\ i = 1, \ldots, M+1\} = \prod_{i=1}^{M+1} \hat{p}_{i,d_i} \tag{6.18}$$

where

$$\hat{p}_{m,d}(t) = \begin{cases} (1 - p_{A,m}(t))y_m(t) & \text{if } d = -1 \\ p_{A,m}(t)y_m(t) + (1 - y_m(t))(1 - p_{A,m}(t)) & \text{if } d = 0 \\ p_{A,m}(t)(1 - y_m(t)) & \text{if } d = 1 \\ 0 & \text{otherwise} \end{cases} \tag{6.19}$$

The probability $p_{A,m}(t)$ that decision maker m chose A is given by

$$p_{A,m}(t) = \frac{\Pr\{u_m(t) = 1\}}{1 + e^{\mu_m(\Delta r(y_m) - \nu_m)}} + \frac{\Pr\{u_m(t) = -1\}}{1 + e^{\mu_m(\Delta r(y_m) + \nu_m)}} + \frac{\Pr\{u_m(t) = 0\}}{1 + e^{\mu_m(\Delta r(y_m))}}. \tag{6.20}$$

$\Pr\{u_m(t) = 1\}$ (respectively, $\Pr\{u_m(t) = -1\}$) is the probability that, among the M decision makers excluding decision maker m, at least $\lceil \frac{M+1}{2} \rceil$ chose A (respectively, B) at time t. $\Pr\{u_m(t) = 0\} = 1 - \Pr\{u_m(t) = 1\} - \Pr\{u_m(t) = -1\}$ is the probability that an equal number of A's and B's were chosen. In case Assumption 9(b) holds instead of Assumption 9(b), then the results hold with $\Delta r(y(t))$ replaced with $f(y(t))$.

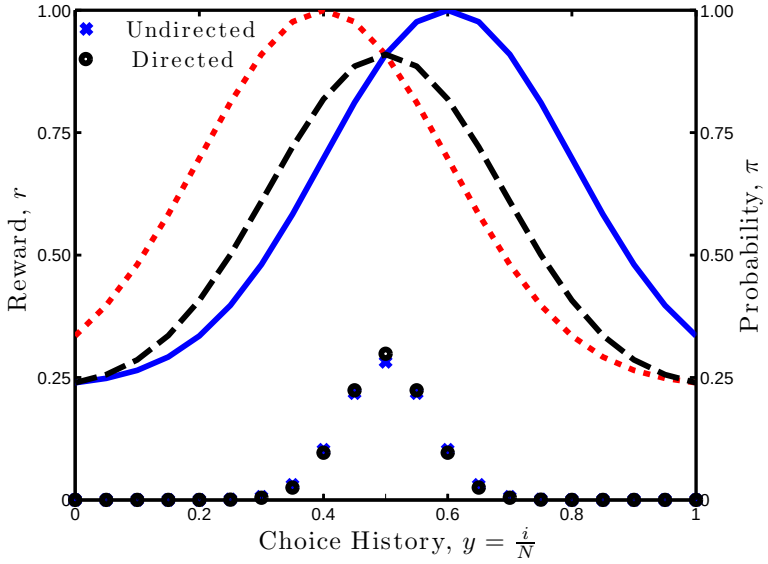
Proof of Proposition 8:

Since for a given choice by decision maker m at time $t+1$, $y_m(t+1)$ can only change from its current value of $y_m(t)$ to $y_m(t)+\frac{1}{N}$, $y_m(t)-\frac{1}{N}$ or stay at $y_m(t)$, we need only compute the probability $\hat{p}_{m,d}$ of each of these three events, $d=1$, $d=-1$, $d=0$, for all $y_m(t) \in \mathcal{Y}$ and each m. Each of these events depends upon the current state $(y_1(t), \ldots, y_M(t))$, as well as each decision maker's most recent choice and oldest choice in their history of N choices, since $y_m(t+1)$ will only differ from $y_m(t)$ for decision maker m if the most recent decision also differs from the oldest decision in the history. The probabilities $\hat{p}_{m,d}$ of (6.19) are derived analogously to (6.1)-(6.3) in Proposition 6 with the probability $p_{A,m}$ that decision maker m chooses A of (6.20) derived analogously to $\bar{p}_A$ of (6.4). The computation of $p_{A,m}(t)$ requires conditioning on the value of $u_m(t)$.

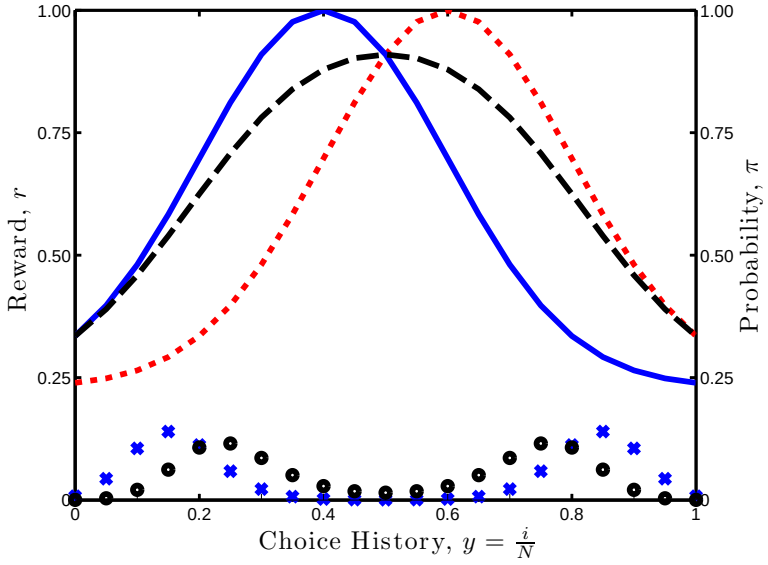
Treating each decision maker's choice as an independent event, the transition probabilities for the group are given by (6.18). Since the probabilities depend only upon $(y_1(t), \ldots, y_M(t))$, the current value of the state at time t, the process is Markov. The case when Assumption 9(b) holds follows similarly. $\square$

Because of the high dimensionality of the matrix $\mathbf{P}$, we compute the expected distributions numerically. This is done by raising $\mathbf{P}$ to a high power so that the elements along each column are equal. Any row in the resulting matrix then has the steady-state distribution as its elements. All rows being equal corresponds to the fact that the probability of transitioning to any of the possible states in the long run is independent of the initial condition.

In Figure 6.4 we show the numerically-computed expected steady-state distribution for one of the decision makers where there is undirected choice feedback and $M=2$. The distribution in the undirected case for the CG task is plotted with x's in Figure 6.4(a). We compare this to the distribution for the focal decision maker in the case of directed choice feedback and $M=2$, which is plotted with circles in Figure 6.4(a) (computed from (6.7)). We see that there is little difference in the



(a)



(b)

Figure 6.4: Comparison of expected steady-state distribution with undirected versus directed choice feedback. In both plots circles correspond to directed feedback (where π is computed from (6.7)) and x's correspond to undirected feedback (where π is computed numerically). (a) CG task with $\mu = 2.6$. (b) DG task with $\mu = 2.9$. In both cases values for μ are chosen in accordance with the fitted values provided in [57].

distributions, suggesting that for our model the CG task results do not depend significantly on whether the feedback is undirected or directed. This is consistent with our comparison between the model predictions in the directed case and the experimental data in the undirected case for the CG task as described in Section 6.3.1.

The same comparison between the undirected case and the directed case for the DG task is shown in Figure 6.4(b). Recall for the DG task that the optimal solution at $y = 0.5$ is a diverging point; because of the symmetry in the reward structure, decision makers diverge to choice sequences with $y > 0.5$ and $y < 0.5$ with equal probability.

The plots in Figure 6.4(b) show that for the DG task the undirected case can differ substantially from the directed case. In this example where $\mu = 2.9$ for all decision makers [57] and $M = 2$, the focal decision maker makes steady-state choices in the undirected case that are further from the optimal solution as compared to the steady-state choices in the directed case. This suggests that undirected feedback reinforces the tendency for decision makers to move toward relatively high and low values of y, leading to reduced performance as compared to the directed feedback case. The result illustrates the influence that the interconnection topology can have on performance of a group of decision makers.

Comparison between directed and undirected topologies is a topic addressed in Chapter 7 where we discuss new experiments with human subjects and various interconnection topologies. The comparison in this section requires numerical computation of distributions for the undirected case. In the numerical computations, the state of each decision maker is included in the Markov chain. The size of the state space prohibits numerical computations for $M > 2$.

This concludes the analysis of this chapter. We are now in a position to apply the predictive tools developed here. In the following chapter, we use our tools to design mixed teams of decision makers.

Chapter 7

Experiments and Design of Decision-Making Dynamics

The success of our approach has allowed us to perform analytic validations of the agreement of the soft-max choice model with experimental data. This has been presented in Chapters 5 and 6. This chapter takes our analysis into a new realm and makes predictions about the behavior of a team of decision makers, some of which receive designed feedback. The findings of the work in this chapter have prompted us to design new experiments that are currently being performed.

The notion of designing social feedback carries along several themes with it, bringing up a number of issues. Studying the change in behavior of a focal decision maker to properties of the social feedback provided is a primary goal of this work. The predictions made in this chapter allow us to find interesting "tipping points" in the decision making. That is, we have isolated types of feedback which, though they differ slightly, can have dramatically different results for a decision maker receiving that feedback.

It is also of interest to develop tools and benchmarks for the design of automated decision makers. By changing properties of the decision-making strategies employed

by members of the team in a principled way, and measuring the effects, we propose tools to develop a methodology for designing decision-making teams.

7.1 Predicted Performance with Designed Decision Makers

In this section we use our predictive model to study performance for the RO task in the social context. Recall that the RO task is considered difficult because the decision maker must endure a period of very poor performance in order to find the optimal choice sequence and even if the optimal solution is found, the RO reward curves make it challenging to sustain. To investigate the influence of choice feedback on the ability of the focal individual to find the optimal solution in the RO task, we design the decision-making parameters of the group of M others who provide feedback to the focal individual and examine the resulting decision-making dynamics using our predictive model. We consider a heterogeneous group of decision makers in the social TAFC task as defined in Section 3.3 of Chapter 3. We illustrate with an example in which a change in the decision-making parameters of only one individual in the group providing feedback can have a dramatic (and positive) effect on performance of the focal decision maker.

Consider a focal decision maker who receives choice feedback from $M = 4$ other decision makers. We implement our design by prescribing the probability $p_{A,m}$ that decision maker m chooses A, for $m = 1, 2, 3, 4$. We consider constant (and heterogeneous) values for $p_{A,m}$. Figure 7.1 shows the expected steady-state distribution for the focal individual with choice feedback from four decision makers with two different designs. In both designs we set $p_{A,1} = .05, p_{A,2} = .95$, and $p_{A,3} = 0.5$; this means that among the four decision makers providing feedback, decision maker 1 sustains choice sequences near the local optimum at $y = 0$, decision maker 2 finds and sustains

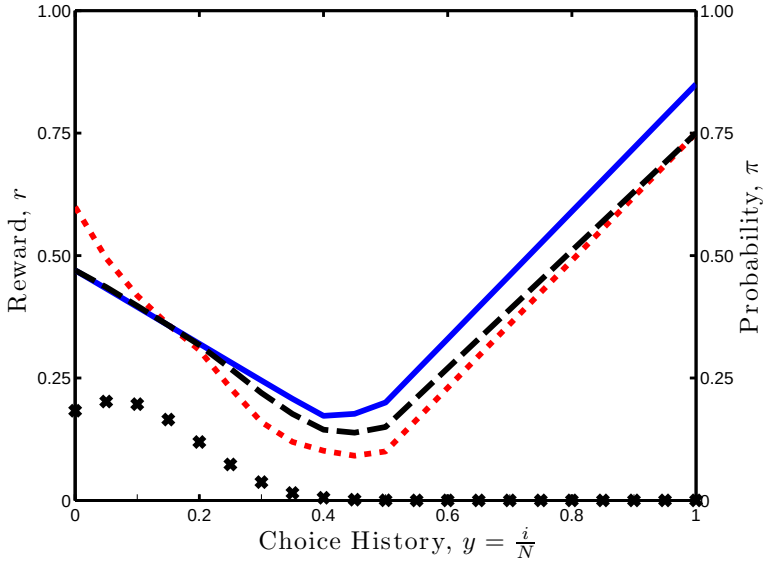
the global optimal solution at $y = 1$ and decision maker 3 is a "noisy" player who randomly chooses A or B. Designing the choice probabilities for these three decision makers in this way does not create a bias towards A or B for the focal decision maker. The difference between the two designs comes in how we prescribe the probability $p_{A,4}$ for decision maker 4. In the first design, corresponding to Figure 7.1(a), we set $p_{A,4} = 0.05$, i.e., decision maker 4 sustains choices near the local optimum at $y = 0$. However, in the second design, corresponding to Figure 7.1(b), we set $p_{A,4} = 0.95$, i.e., decision maker 4 sustains choices near the global optimum at $y = 1$.

The results of the steady-state distribution show that in the first design, the focal decision maker does not find the global optimal but rather chooses B most of the time and remains close to the local optimal solution at $y = 0$ (Figure 7.1(a)). However, the results of the second design are dramatically different. The change in the probability of choosing A by decision maker 4 in the second design makes all the difference in helping the focal decision maker find and sustain choices close to the global optimal solution (Figure 7.1(b)).

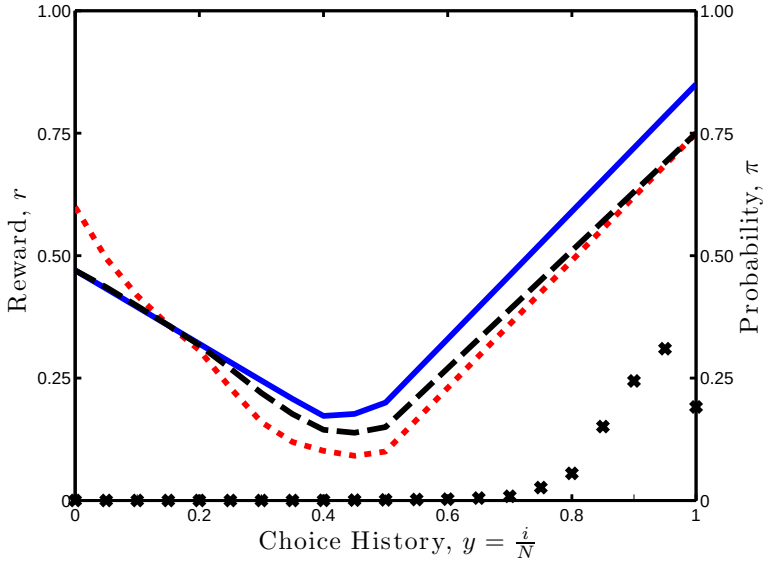
The space of possible designed feedback is large and we want to make predictions in a principled way. In the RO task, a measure of good performance is whether decision makers "discover" the optimum choice sequence; i.e. find the maximum value in the reward structure. By calculating the probability that a decision maker's choice history will have proportion of A greater than a chosen value, we can predict the likelihood that the optimum choice sequence is discovered. We calculate this probability from the following:

$$\Pr\{y(t) > y_c = \frac{i_c}{N}\} = \sum_{i=i_c}^{N+1} \pi_i(\mu, \nu, \Delta r, p_{A,1}, \ldots, p_{A,4}) \tag{7.1}$$

where y_c is a critical value of y that we pick and i_c is the corresponding number of choices A made in the last N trials to achieve y_c. Recall Equation 6.7 where π_i is expressed analytically.



(a)



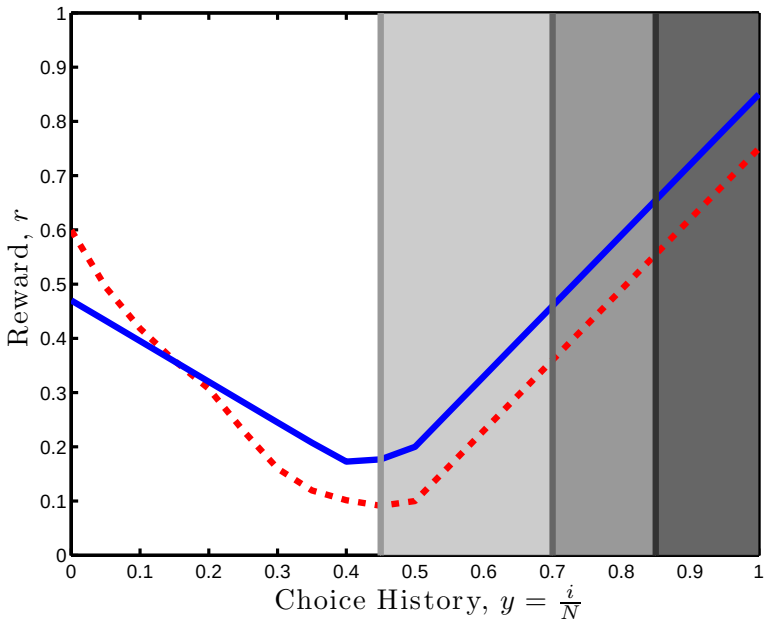
(b)

Figure 7.1: Expected steady-state distribution of $y(t)$ for a focal decision maker receiving choice feedback from four designed decision makers in the RO task. In each case decision makers 1-3 choose A with probability $p_{A,1} = .05$, $p_{A,2} = .95$, and $p_{A,3} = 0.5$. (a) Decision maker 4 chooses A with probability $p_{A,4} = .05$. The distribution for the focal decision maker is shown with circles; the focal decision maker spends most time near the local optimal solution at $y = 0$. (b) Decision maker 4 chooses A with probability $p_{A,4} = .95$. The distribution for the focal decision maker is shown with x's; the focal decision maker finds and sustains choices near the global optimal solution at $y = 1$.

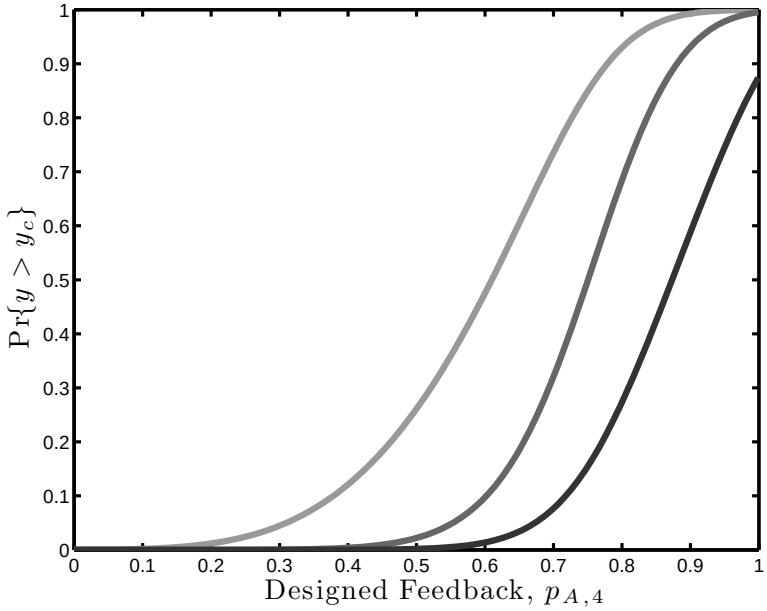
We now have an analytic function which depends on feedback from the designed decision maker 4, and also parameters of the focal decision maker and environment. This probability function is plotted for three cases of y_c in Figure 7.2(b). The values of y_c are depicted in Figure 7.2(a) by the vertical lines. Consider $y_c = 0.45$, plotted in light grey. Should a decision maker make choice sequences such that $y(t) > 0.45$, then they will reside in the domain lying to the right of the light gray vertical line in Figure 7.2(a) (which is shaded in light grey). In Figure 7.2(b) the corresponding probability of an individual converging to such choice sequences is plotted as a function of $p_{A,4}$. For low $p_{A,4}$ we have that $\Pr\{y(t) > 0.45\} = 0$. There is a critical value of $p_{A,4}$ roughly equal to 0.6, where half the time is spent with $y(t) > 0.45$. For increasing $p_{A,4}$ the decision maker is influenced to have higher proportion of A in their history, moving closer to the optimal. As we increase y_c greater $p_{A,4}$ is required to restrict the behavior to domains corresponding to the higher reward. Figure 7.2 shows the prediction for $y_c = 0.70$ (medium grey) and $y_c = 0.85$ (dark grey).

The case of $y_c = 0.85$ defines a domain which is essentially optimal. Should choices be made such that $y(t) > 0.85$ for a sustained period, much higher rewards will be earned in the task. We see from the dark grey curve in Figure 7.2(b) that for $p_{A,4} = 0.90$, the predicted behavior is to spend more than half the time very close to the optimal solution. We have chosen a value of $p_{A,4} = 0.95$ in experiments and are expecting subsequent choices sequences to be near optimal for the majority of the experiment's duration as shown in Figure 7.1(b).

We saw in Section 6.3 that for the easy CG task, choice feedback was detrimental to performance. It is significant that for the difficult RO task our model predicts that performance can be significantly improved by choice feedback. Such results may prove useful in the design of human-robot decision-making teams where well justified methodology is needed for programming the decision-making parameters of



(a)



(b)

Figure 7.2: Probability that a focal decision maker receiving choice feedback will have choice sequences such that $y(t) > y_c$ in the RO task. (a) Domains of interest for three cases of y_c. The vertical lines coincide with $y_c = 0.45$, 0.70, and 0.90 (light, medium, and dark grey, respectively) and corresponding domains shaded. (b) The probability, given by (7.1) is plotted as a function of the designed feedback $p_{A,4}$ under three cases: $y_c = 0.45$ (light grey), $y_c = 0.70$ (medium grey), $y_c = 0.85$ (dark grey).

the robots to optimize team performance. In the following section we detail the experiments planned to test these predictions.

7.2 Designed Feedback Experiments

Experiments that will test the hypothesis of this chapter are currently underway. The experiments, led by Damon Tomlin of the Psychology Department, aim to test human subjects receiving choice feedback from designed choice sequences to experimentally verify our prediction that such a change in the choice feedback may result in a significant performance increase for a focal subject.

Our predictions will be tested by each of the canonical reward structures we consider for the TAFC task. While we expect interesting findings in each of the reward structures, and aim to compare our directed feedback analysis of Chapter 6 with experimental results, the potential for increasing performance in the RO reward structure is of particular interest.

The behavioral experiments which are currently underway have two components. Since the social feedback analysis of this thesis relies on assuming feedback is directed, we test a condition with directed social feedback in these new experiments. We also switch the social feedback to designed feedback from decision makers that we program. These can be interpreted as robotic decision makers that interact with a human subject as a peer.

- Directed Social Feedback: Subjects in this experiment will face the TAFC task under choice feedback conditions, but rather than receive feedback from $M = 4$ other subjects making live choices in real time (as was the case in previous experiments), they will be receiving feedback from $M = 4$ "typical" decision makers who had already participated in an experiment without social feedback.

This allows us to acquire data for a focal decision maker in a network with directed choice feedback.

- Designed Social Feedback: Subjects in this experiment will face the TAFC task with directed choice feedback, but the feedback they receive will come from $M = 4$ *designed decision makers*. The design of the decision making dynamics for the four automated decision makers in the group will be as discussed in Section 7.1. This experiment allows us to test our predictive capability, in the RO task and other reward structures and help us to explore the effects of designed feedback in each scenario.

The two primary goals are to perform experiments with directed feedback from actual human subjects, and also to study the effects of designed social feedback. The latter goal is one that should require extensive experimentation. While our analysis in this chapter sheds light on a potentially important effect which arises in the RO structure, determining the broad-reaching effects of designing feedback in teams of decision makers is a sizable challenge. For instance, the effect on variance, exploration, engagement (boredom), and even trust, are all critical issues which should be considered.

Chapter 8

Conclusion and Ongoing Work

The contributions of this thesis are comprised of several model-based analyses which have led to the development of predictive tools to assist the design of mixed teams. Predictions we've made using tools developed in this work have prompted the design of new experiments. In the experiments discussed in Chapter 7, we aim to test our hypotheses and gather data for mixed teams of human/robot decision makers in the TAFC task. Relevant applications and methods of interaction between humans and robots are also of particular interest. Through developing a robotic testbed that supports real-time operation of multiple, robotic vehicles in a three-dimensional field, we have directed our research toward more natural and relevant tasks for human and robotic decision makers. Several experiments are planned that will use the multi-vehicle testbed. Plans for those experiments are described in this chapter.

8.1 Summary

In the work of this thesis we have focused primarily on the TAFC task as presented in Chapter 3. In Chapters 4 through 6, we presented the tools we developed to provide a predictive capability for decision making in teams with social feedback. Chapters 4 and 5 focus on an individual decision maker that does not receive social feedback.

In Chapter 6 our analysis is extended to allow for social feedback. In Chapter 7 we used our results to identify a scenario where robotic decision makers designed into a mixed team with humans can significantly improve performance.

While these tools apply directly to a subset of scenarios that the TAFC task can be mapped to, the principles that we uncover in this work are likely to extend to different and more complex tasks. In this chapter, we present plans to study a spatial decision-making task with physical robots that work together with humans. Details of the spatial task and plans for experiments with out robotic testbed appear in Section 8.2.2. There are also many real-world scenarios that the TAFC task can be mapped to. In Section 8.2.1 we motivate an oil spill application for the TAFC task.

In Chapter 2 we presented details of the robotic testbed which is comprised of a fleet of underwater vehicles that navigate a three-dimensional, virtual (imposed) resource field. The testbed will be used as part of a study to investigate decision-making with humans and physical robots working together. Some planned experiments are detailed in Section 8.2.2. Experiments that focus specifically on joint decision making in mixed teams, but without using physical robots, were presented in Chapter 7.

8.2 Ongoing and proposed work

In this section we lay out experiments that are designed to incorporate human decision makers in an integrated robotic decision-making task. The goal is to perform experiments with physical robots that interact in real time with human decision makers and to study a task in which the alternatives (choices) are given concrete meaning. In Section 8.2.1 we introduce a relevant application for the TAFC task that we study in this thesis. Through considering that application, and developing new experimental paradigms with our psychologist and neuroscientist collaborators, we have created a new task which more naturally maps to scenarios that have a spatial element. That

task is described in Section 8.2.2. Experiments with our robotic testbed will test the ability of humans and robots to work together in a relevant scenario such as this. The experiments use the hardware described in Chapter 2 and are laid out in Section 8.2.3.

8.2.1 Application

Consider a decision-making application in which a team of decision makers works together in a TAFC task. Take, for example, the aftermath of a large oil spill in the ocean. A cleanup operation is taking place, and equipment is being distributed in the field to collect the spilled oil. To maximize efficiency and speed the cleanup process, a strategy that sends oil-collecting vessels to the areas of maximal concentration of oil is employed. The problem at hand is then an information collection problem; i.e. the decision-making team must determine the areas of maximal concentration of oil.

The team has access to information from autonomous vehicles equipped with sensors that are also deployed in the contaminated area. Consider two autonomous vehicles that move around in a fixed region. The vehicles have automated control that prescribes their motion in the field. The movement is regular, although each vehicle moves in a different pattern, perhaps according to a different control algorithm (each may be optimized for different behavior). One vehicle may be programmed to find the centroid of an area with given concentration, another may be programmed to track gradients, or find boundaries in the field.

The decision-making team must acquire information from the autonomous vehicles deployed in the oil spill region. The data is acquired in real time, and informs the decision of where to send oil-collecting vessels. An operator in the team may query a vehicle to get an assessment at that point. Consider the case in which the cost of communication is high (which is certainly true for systems that operate below the ocean surface). The operator then must choose between one of the vehicles to query.

We denote the vehicles A and B. The metric (or "reward") received in response to a query should be one indicating the impact that the corresponding measurement and location have on the assessment. The goal for the team is to maximize that impact.

Vehicle dynamics, and also the dynamics of the oil in the ocean, make for a time-varying environment and a complex resource allocation and decision-making problem. Although we have described a scenario that maps to the TAFC task we study, it is likely the case that rewards received by human supervisors in this application are noisy and much more complex than those studied in our experiments or the analyses of Chapters 4-7. There are almost certainly, however, likely to be characteristics in the reward structure which we have studied in this work, and subsequent effects on decision-making behavior should be expected. For instance, convergent matching points, and hidden points of optimality should exist for some time periods in this task. It should be true that certain choice sequences would lead to higher performance.

Should this exercise be one that continues over time, there should be correlations in the distribution of the oil since its movement follows some prescribed dynamics. In that case, there is something to be learned about the environment and social feedback should prove valuable. The approach we described in Chapter 7 could be of considerable use. We envision tasking mixed teams with decision making in applications of this nature. Those teams could be made up of some human and some automated decision makers.

8.2.2 Experiments with Human-Robot Teams

We aim to gather experimental data for human subjects that work jointly with robotic platforms in a real-world setting. Our robotic testbed allows us to validate results and test hypotheses in a controlled environment which can be quickly reconfigured for an array of experiments. This section details plans that have been made to study inte-

grated decision making with humans and robots exploring three-dimensional resource fields using the testbed presented in Chapter 2.

We have designed a number of experiments to run with humans and robots making joint decisions about where and how to move in the field, and how to share information within the team. While at first we are running experiments that integrate a decision maker with a role equivalent to that of subjects in the two-alternative, forced-choice task, we are also taking what we have learned from our predictive capability and designing more complex paradigms for integrating the human input.

The bulk of experiments planned for the future of the testbed discussed in Chapter 2 will emphasize integrating human decision making for a spatial task. The spatial task is similar to the TAFC task. In fact, the task we consider grew out of the studies, and is informed by findings with the TAFC task. While the TAFC task can be thought of as one-dimensional (the state of decision makers is confined to an interval defined by $y \in [0, 1]$) the spatial task is two dimensional. The action space (available choice alternatives) for the spatial task is larger as well.

In the spatial task developed by Damon Tomlin, subjects see a grid of locations that they can visit. They make the choice to move from their current location to another location by making one step either up, down, left, or right. They may also choose to remain at the current location. The size of the action space is then 5 alternatives.

Damon Tomlin has run some experiments with human subjects performing the spatial task. An analysis of the data that was performed by Paul Reverdy and Andrea Nedic has indicated that subjects in the spatial task have a point in the space intended as a destination. The hypothesis is that individual choices are made to navigate toward a target, rather than for the sake of visiting each point along the way. This preliminary finding has motivated a slight change to the spatial task from its original design by Damon Tomlin.

One convenient, and common, way to task autonomous vehicles that are deployed in a field is to assign a target to each vehicle. Vehicles have onboard control that directs them to the assigned target. In this context, vehicles are sequentially assigned targets, once they arrive at a target they remain on location or continue to the next target. By modifying the spatial task to allow decision makers to choose targets, rather than the five navigation-based choices (up/down, left/right, and stay), we can link rigorous studies of human decision making in an abstract task with a concrete problem of interaction and joint decision making with humans and robots.

8.2.3 Experimental variations

We have a plan to go about studying joint decision making with humans and robots using our multi-vehicle, robotic testbed and a remote human interface (see Chapter 2 for hardware details). The approach is to use physical robots and to add complexity to the study in a systematic way. The ultimate goal is to draw conclusions about teams of human decision makers who are integrated with a team of robotic decision makers. The robotic elements include the mobile robots that carry out tasks and explore environments, but may also include automated decision-making capabilities which act as peers to the human decision makers. This concept of mixed teams is discussed in depth in Chapter 6.

A subject in the robotic testbed experiments interacts with the remote interface we have designed and built to incorporate human input. The human is provided with live video feeds from the robots and a map with vehicle location(s) is displayed as well. The map is a two-dimensional area that shows vehicle location(s) in the horizontal plane. The interface allows for the subject to select a location on the map, thereby identifying that location as a target. Feedback on performance is also displayed. Performance in these experiments depends upon the underlying reward structure which is imposed via a virtual resource field. Several resource fields will be

imposed, each being mapped from reward structures previously studied in the spatial task. In each variation the duration of the experiment is fixed.

The first variation of our experiment is designed to include one human decision maker in the loop. The subject chooses targets in a two-dimensional area with an imposed resource field. A single vehicle moves to the assigned target (a cylinder about the point in three-dimensions). Then the vehicle remains on station until a new target is selected. Targets may also be switched before a vehicle arrives. Feedback of the reward (resource measured by the robot along its route to the target) is provided at a constant, steady rate after each target is chosen. The subject continues with the task, sequentially identifying targets, or choosing to keep the most recent target active. Their goal is to maximize accumulated reward. In this first variation, the imposed resource field is time invariant, and two-dimensional.

A number of variations of the task can be considered in the case of a single human subject. These include allowing for different reward feedback paradigms, and making use of multiple vehicles. In each variation, feedback of the reward is provided at a constant, steady rate, but can also include additional information. For example, a map of the estimated field based on accumulated measurements or a map of uncertainty in the information can be displayed by the interface using an overlay so that vehicle and target locations are displayed together with the additional information. When uncertainty in the information is provided as a performance metric, the subject's goal is to minimize the uncertainty.

In experiments involving multiple vehicles, several methods of deployment are planned. The subject is still tasked with selecting targets, and vehicles are used to visit those targets, but now a coordinated formation of vehicles can collect multiple measurements and provide filtered, more accurate estimates of the field at the formation's centroid. Vehicles can travel toward the selected target while also climbing the gradient of the imposed field, or performing other adaptive coordinated behav-

iors such as dynamically servicing targets to minimize wait times. We also plan to add constraints to the physical space; i.e. obstacles, danger zones, adversaries, and additional environmental dynamics such as current.

Complexity can also be added to the imposed resource field. We plan to use fields that are time-varying and "consumable" so that measurements depend on the time and location they are taken. Ultimately, fields considered in these experiments will also vary in all three spatial dimensions.

The next step is to incorporate input from multiple subjects, each selecting targets in parallel. This will require additional infrastructure to accommodate multiple human decision makers. The interface in its current form is designed to incorporate a single human in the system. Multiple subjects will choose targets in a two-dimensional landscape (in parallel). Each of the subjects' targets is serviced by a vehicle (or a vehicle group). We also consider the case of mixed human / robot decision makers in this context. Feedback of the reward can be provided with the same variations mentioned previously but will also include social feedback so that information is shared. One practical way of sharing relevant information is to provide an estimate of the resource field as computed from all of the measurements (from each subject).

There is a large space of alternative variations of these experiments. It should be noted that another way to integrate a team with this testbed is by enforcing the constraint that only one decision is made for the group. This is in contrast to the tasks we have studied since each decision maker in the group makes a decision in parallel to the rest of the group. Instead, all of the decision makers can work together to agree on single, collective decision. The testbed can be adapted to incorporate this type of decision-making protocol as well. It is likely that such studies would prove useful and result in interesting findings.

By starting with a well-defined task that can be modeled and studied analytically, we have set about building an experimental study of joint decision making in mixed

teams that has a strong foundation in well-understood principles. Much work is yet to be done toward this end, but the work in this thesis provides a starting point for designing mixed teams of decision makers that are comprised of humans and robots. Though the TAFC task is simple in its formation, we've shown that human subjects exhibit suboptimal behavior when faced with some reward structures and are also sensitive to social feedback – findings that suggest formal studies of the decision making in this context are warranted.

Through systematic increases in the complexity of our analysis, and collaborations with psychologists and neuroscientists performing experiments with human subjects, we've developed valuable tools for predicting decision-making behavior in mixed teams that share social feedback. In this section we've outlined how to move forward with these studies with the goal of further developing design tools for building mixed teams. These studies prove to be rewarding from both a scientific and an engineering perspective. As this research continues, advancements will be made not only toward understanding how humans perform complex tasks, but also toward how to engineer systems to assist humans when the need arises.

Bibliography

[1] R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. Task planning for human-robot interaction. In *Proc. Joint Conf. on Smart Objects and Ambient Intelligence*. ACM, 2005.

[2] B. Allen, R. Stokey, T. Austin, N. Forrester, R. Goldsborough, M. Purcell, and C. von Alt. Remus: a small, low cost auv; system description, field trials and performance results. In *OCEANS '97. MTS/IEEE Conference Proceedings*, volume 2, pages 994 –1000 vol.2, oct 1997.

[3] B. Anderson and J. Crowell. Workhorse auv – a cost-sensible new autonomous underwater vehicle for surveys/ soundings, search & rescue, and research. In *OCEANS, 2005. Proceedings of MTS/IEEE*, pages 1 –6, sept. 2005.

[4] J. D. Anderson and J. Anderson. *Introduction to Flight*. McGraw-Hill, New York, NY, 5 edition, March 2004.

[5] R. C. Arkin. *Behavior-Based Robotics (Intelligent Robotics and Autonomous Agents)*. MIT Press, 1999.

[6] R. Bachmayer, N. E. Leonard, J. Graver, E. Fiorelli, P. Bhatta, and D. Paley. Underwater gliders: recent developments and future applications. Taipei, Taiwan, 2004.

[7] R. Bachmayer and N.E. Leonard. Experimental test-bed for multi-vehicle bontrol, navigations and communication. In *Proceedings of the 12th International Symposium on Unmanned Untethered Submersible Technology*, Durham NH, 2001.

[8] D. Baronov and J. Baillieul. Reactive exploration through following isolines in a potential field. In *Proc. of 2007 American Control Conference*, pages 2141–2146, 2007.

[9] P. Bhatta. *Nonlinear Stability and Control of Gliding Vehicles*. PhD thesis, Princeton University, Mechanical and Aerospace Engineering designation 3159T, 2006.

[10] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113:700–765, 2006.

[11] R. Bogacz, S. M. McClure, J. Li, J. D. Cohen, and P. R. Montague. Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, 1153:111–121, 2007.

[12] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59:119–155, July 2003.

[13] C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4):167 – 175, 2003.

[14] B. Brehmer. Dynamic decision making: Human control of complex systems. *Acta Psychologica*, 81(3):211 – 241, 1992.

[15] A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on*, volume 4, pages 4138 – 4142 vol.4, 2002.

[16] D. J. Bruemmer, D.A. Few, R. L. Boring, J.L. Marble, M.C. Walton, and C. W. Nielsen. Shared understanding for collaborative control. In *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, volume 35(4), pages 494–504, 2005.

[17] C. Caicedo-Nunez and N.E. Leonard. Symmetric coverage of dynamic mapping error for mobile sensor networks. In *American Control Conference (ACC), 2011*, pages 4661 –4666, 29 2011-july 1 2011.

[18] M. Cao, A. Stewart, and N. E. Leonard. Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis. In *Proc. of the 47th IEEE Conference on Decision and Control*, pages 1127–1132, Cancun, Mexico, 2008.

[19] M. Cao, A. Stewart, and N. E. Leonard. Convergence in human decision-making dynamics. *Systems and Control Letters*, 59:87–97, 2010.

[20] J. Casper and R.R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 33(3):367 – 385, june 2003.

[21] K. Cleary and C. Nguyen. State of the art in surgical robotics: Clinical applications and technology challenges. *Computer Aided Surgery*, 6(6):312–328, 2001.

[22] A. Clodic, R. Alami, V. Montreuil, S. Li, B. Wrede, and A. Swadzba. A study of interaction between dialog and decision for human-robot collaborative task achievement. In *Robot and Human interactive Communication*, pages 913–918, 2007.

[23] M.L. Cummings and P.J. Mitchell. Predicting controller capacity in supervisory control of multiple uavs. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(2):451 –460, march 2008.

[24] S. H. Dandach, R. Carli, and F. Bullo. Accuracy and decision time for sequential decision aggregation. August 2010. To appear.

[25] P. Dario, B. Hannaford, and A. Menciassi. Smart surgical tools and augmenting devices. *Robotics and Automation, IEEE Transactions on*, 19(5):782 – 792, oct. 2003.

[26] B. Donmez, M. L. Cummings, and H. D. Graham. Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics*, 51(5):718–729, October 2009.

[27] M. P. Eckstein, K. Das, B. T. Pham, M. F. Peterson, C. K. Abbey, J. L. Sy, and B. Giesbrecht. Neural decoding of collective wisdom with multi-brain computing. *NeuroImage*, In Press, Corrected Proof:–, 2011.

[28] D. M. Egelman, C. Person, and P. R. Montague. A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, 10:623–630, 1998.

[29] C.C. Eriksen, T.J. Osse, R.D. Light, T. Wen, T.W. Lehman, P.L. Sabin, J.W. Ballard, and A.M. Chiodi. Seaglider: a long-range autonomous underwater vehicle for oceanographic research. *Oceanic Engineering, IEEE Journal of*, 26(4):424 –436, oct 2001.

[30] B. Etkin. *Dynamics of Flight*. Wiley, New York, 1959.

[31] E. Fiorelli. *Cooperative Vehicle Control, Feature Tracking and Ocean Sampling*. PhD thesis, Princeton University, Mechanical and Aerospace Engineering, 2005.

[32] E. Fiorelli, N. E. Leonard, P. Bhatta, D. Paley, R. Bachmayer, and D.M. Fratantoni. Multi-auv control and adaptive sampling in Monterey Bay. In *IEEE Journal of Oceanic Engineering*, volume 31, pages 935–948, 2006.

[33] E. Fiorelli, N.E. Leonard, P. Bhatta, D. Paley, R. Bachmayer, and D.M. Fratantoni. Multi-auv control and adaptive sampling in Monterey Bay. In *Autonomous Underwater Vehicles, 2004 IEEE/OES*, pages 134 – 147, june 2004.

[34] T. Fong, C. Kunz, L.M. Hiatt, and M. Bugajska. The human-robot interaction operating system. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 41–48, New York, NY, USA, 2006.

[35] T. Fong, I. Nourbakshsh, C. Kunz, L. Fluckiger, J. Schreiner R. Ambrose, R. Burridge, R. Simmons, L. M. Hiatt, A. Schultz, J. G. Trafton, M. Bugajska, and J. Scholtz. The peer-to-peer human-robot interaction project. In *AIAA Space*, pages 2005–6750, 2005.

[36] T. Fong and C. Thorpe. Vehicle teleoperation interfaces. *Autonomous Robots*, 11:9–18, 2001. 10.1023/A:1011295826834.

[37] T. I. Fossen. *Guidance and Control of Ocean Vehicles*. John Wiley and Sons, New York, 1994.

[38] J. Graver. *Underwater Gliders: Dynamics, Control and Design*. PhD thesis, Princeton University, Mechanical and Aerospace Engineering, 2005.

[39] T.M. Gureckis and B.C. Love. Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, 53:180–193, 2008.

[40] R. Herrnstein. Rational choice theory: necessary but not sufficient. *American Psychologist*, 45:356–367, 1990.

[41] R. Herrnstein. *The Matching Law: Papers in Psychology and Economics*. Harvard University Press, Cambridge, MA, USA, 1997. Edited by Howard Rachlin and David I. Laibson.

[42] D. Kahneman and A. Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, 1979.

[43] N.J. Kasdin and D. A. Paley. *Engineering Dynamics*. Princeton University Press, 2011.

[44] T. Kaupp and A. Makarenko. Measuring human-robot team effectiveness to determine an appropriate autonomy level. In *IEEE International Conference on Robotics and Automation*, pages 2146–2151, Pasadena, CA, 2008.

[45] T. Kaupp, A. Makarenko, and H. Durrant-Whyte. Human-robot communication for collaborative decision making – a probabilistic approach. *Robotics and Autonomous Systems*, 58(5):444 – 456, 2010.

[46] Y Kim, W. Yoon, H. Kwon, Y. Yoon, and H. Kim. A cognitive approach to enhancing human-robot interaction for service robots. In *Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design*, volume 4557, pages 858–867. Springer Berlin / Heidelberg, 2007.

[47] A. Kron, G. Schmidt, B. Petzold, M.I. Zah, P. Hinterseer, and E. Steinbach. Disposal of explosive ordnances by use of a bimanual haptic telepresence system. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 2, pages 1968 – 1973 Vol.2, 26-may 1, 2004.

[48] H. Lamb. *Hydrodynamics*. Cambridge University Press, London, 1932.

[49] N. E. Leonard, D. Paley, R. Davis, D.M. Fratantoni, F. Lekien, and F. Zhang. Coordinated control of an underwater glider fleet in an adaptive ocean sampling field experiment in Monterey Bay. In *Journal of Field Robotics*, volume 27, pages 718–740, 2010.

[50] Merlin Systems Corp. Limited. Introduction to the Miabot Pro Version 1. `http://www.merlinrobotics.co.uk/`.

[51] M. Maurette. Mars rover autonomous navigation. *Autonomous Robots*, 14:199–208, 2003. 10.1023/A:1022283719900.

[52] B. W. McCormick. *Aerodynamics, Aeronautics and Flight Mechanics*. Wiley, New York, 1979.

[53] P. R. Montague and G. S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36:265–284, 2002.

[54] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16:1936–1947, 1996.

[55] R.R. Murphy. Human-robot interaction in rescue robotics. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(2):138–153, may 2004.

[56] A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice, and J.D. Cohen. A simple decision task in a social context: experiments, a model, and preliminary analyses of behavioral data. In *Proc. of the 47th IEEE Conference on Decision and Control*, pages 1115–1120, Cancun, Mexico, 2008.

[57] A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice, and J.D. Cohen. A decision task in a social context: Behavioral experiments, models, and analyses of behavioral data. In *Proceedings of the IEEE*, 2011. To appear.

[58] M. Nowak and K. Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364:56–58, 1993.

[59] B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, Berlin, 2003.

[60] D. A. Paley. *Cooperative Control of Collective Motion for Ocean Sampling with Autonomous Vehicles*. PhD thesis, Princeton University, Mechanical and Aerospace Engineering designation 3172T, 2007.

[61] B. Parsons and J. Preston. The beluga project: Development of a testbed for autonomous underwater vehicles. Undergraduate thesis, Princeton University, Mechanical and Aerospace Engineering, 2011.

[62] D. Raghunathan and J. Baillieul. Search decisions in a game of polynomial root counting. In *Proceedings of the American Control Conference*, pages 2396–2403, Baltimore, MD, 2010.

[63] R. Ratcliffe. A theory of memory retrieval. In *Psychol. Rev.*, volume 85, pages 59–108, 1978.

[64] R. Ratcliffe and J. N. Rouder. A diffusion model account of masking in two-choice letter identifification. In *Journal of Experimental Psychology*, volume 26, pages 127–140, 2000.

[65] R. Ratcliffe, T. Van Zandth, and G. McKoon. Connectionist and diffusion models of reaction time. In *Psychol. Rev.*, volume 106(2), pages 261–300, 1999.

[66] K. Savla, T. Temple, and E. Frazzoli. Human-in-the-loop vehicle routing policies for dynamic environments. In *Proc. of the 47th IEEE Conference on Decision and Control*, 2008.

[67] P. Scerri, D. V. Pynadath, and M. Tambe. Towards adjustable autonomy for the real world. In *Journal of Artificial Intelligence Research*, volume 17, pages 171–228, 2002.

[68] M. Schoendorf. Robotic fish. Undergraduate thesis, Princeton University, Electrical Engineering Department, Thesis number 24652, 2010.

[69] J. Scholtz. Theory and evaluation of human robot interactions. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, page 10 pp., jan. 2003.

[70] N. Sharkey. Computer science: The ethical frontiers of robotics. In *Science*, volume 322, pages 1800–1801, 2008.

[71] J. Sherman, R.E. Davis, W.B. Owens, and J. Valdes. The autonomous underwater glider "spray". *Oceanic Engineering, IEEE Journal of*, 26(4):437 –446, oct 2001.

[72] P. Simen and J. D. Cohen. Explicit melioration by a neural diffusion model. *Brain Research*, 1299:99–117, 2009. Submitted to *Brain Research*.

[73] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, March 1956.

[74] P. L. Smith and R. Ratcliff. Psychology and neurobiology of simple decisions. In *Trends in Neuroscience*, volume 27(3), pages 261–300, 2004.

[75] R. D. Sorkin, R. West, and D. E. Robinson. Group performance depends on the majority rule. In *Psychological Science*, volume 9, pages 456–463, 1998.

[76] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, HRI '06, pages 33–40, New York, NY, USA, 2006. ACM.

[77] R. F. Stengel and J. R. Broussard. Prediction of pilot-aircraft stability boundaries and performance contours. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume SMC-8, NO. 5, 1978.

[78] A. Stewart, M. Cao, and N. E. Leonard. Steady-state distributions for human decisions in two-alternative choice tasks. In *Proc. of the American Control Conference*, number 2378-2383, Baltimore, MD, 2010.

[79] A. Stewart, M. Cao, A. Nedic, D. Tomlin, and N. E. Leonard. Towards human-robot teams: Model-based analysis of human decision making in tow alternative choice tasks with social feedback. In *Proceedings of the IEEE*, 2011. To appear.

[80] A. Stewart and N. E. Leonard. The role of social feedback in steady-state performance of human decision making for two-alternative choice tasks. In *Proc. of the 49th IEEE Conference on Decision and Control*, number 3796-3801, Atlanta, GA, 2010.

[81] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.

[82] D. Swain, I. D. Couzin, and N.E. Leonard. Real-time feedback-controlled robotic fish for behavioral experiments with fish schools. In *Proceedings of the IEEE*, 2011.

[83] H. M. Taylor and S. Karlin. *An Introduction to Stochastic Modeling -3rd ed.* Academic Press, 1998.

[84] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 35(4):460–470, 2005.

[85] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974.

[86] K. P. Valavanis. *Advancements in Unmanned Aerial Vehicles: State of the Art and the Road to Autonomy*. Springer, 2007.

[87] L. Vu and K. Morgansen. Modeling and analysis of dynamic decision making in sequential two-choice tasks. In *Proc. of the 47th IEEE Conference on Decision and Control*, 2008.

[88] J. P. Wangermann and R. F. Stengel. Principled negotiation between intelligent agents: a model for air traffic management. *Artificial Intelligence in Engineering*, 12(3):177 – 187, 1998.

[89] R. Washington, K. Golden, J. Bresina, D.E. Smith, C. Anderson, and T. Smith. Autonomous rovers for mars exploration. In *Aerospace Conference, 1999. Proceedings. 1999 IEEE*, volume 1, pages 237 –251 vol.1, 1999.

[90] D.C. Webb, P.J. Simonetti, and C.P. Jones. Slocum: an underwater glider propelled by environmental energy. *Oceanic Engineering, IEEE Journal of*, 26(4):447 –452, oct 2001.

[91] D. A. Wiegman and S. A. Shappell. *A human error approach to aviation accident analysis: The human factors analysis and classification system.* Ashgate Publishing Limited, Burlington, VT, 2003.

[92] C. Woolsey. *Engergy Shaping and Dissipation: Underwater Vehicle Stabilization Using Internal Rotors.* PhD thesis, Princeton University, Mechanical and Aerospace Engineering, 2001.

[93] H.A. Yanco and J. Drury. Classifying human-robot interaction: an updated taxonomy. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 3, pages 2841 – 2846 vol.3, oct. 2004.

[94] F. Zhang, D. M. Fratantoni, D.A. Paley, J. Lund, and N.E. Leonard. Control of coordinated patterns for ocean sampling. In *International Journal of Control, special issue on Navigation, Guidance of Uninhabited Underwater Vehicles*, volume 80, pages 1186–1199, July 2007.