# Precision Analytics: Learning and Optimization in the Personalized Setting

By

Yonatan Mintz

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Anil Aswani, Chair
Executive Associate Dean Philip Kaminsky
Professor Peter Bartlett

Spring 2018

Abstract

Precision Analytics: Learning and Optimization in the Personalized Setting

by

Yonatan Mintz

Doctor of Philosophy in Industrial Engineering and Operations Research

University of California, Berkeley

Assistant Professor Anil Aswani, Chair

The recent increase in data and computing resource availability has made the use of data analytics more practical in practice than ever before. In particular, the ubiquity of technologies such as smartphones, wearable devices, and smart sensors has allowed for the collection of a large amount of individual level data. In contrast to traditional data analytics, which relies on wide sampling from a population to draw conclusions (i.e. cross-sectional sampling), this so called deep sampled data can be used through precision analytics to create customized experiences that better serve individuals and organizations. In this thesis, we explore this precision analytics framework that builds upon the fields of reinforcement learning and data driven decision making by extending their results to applications with individual level deep sampled data. One of the main applications of this framework is to systems where a decision maker is interested in applying an intervention (or policy) to affect the behavior of an individual agent or group of agents. Two key challenges that arise when analyzing such systems are that the decision maker may have scarce resources or high risk decisions that constrain how they apply their intervention, and that the decision maker may only have partial knowledge of how an agent will react to the intervention. The main focus of this thesis is to begin to analyze these challenges by providing predictive models that can accurately capture individual agent behavior, new estimation and machine learning techniques to efficiently estimate model parameters, and effective online and batch optimization methods to calculate these interventions. We will also discuss how these approaches can be implemented in practice, particularly in the precision healthcare setting.

*To Babs and Lori...*

# Contents

# List of Figures

# List of Tables

# Acknowledgements

It has been a great pleasure and honor to work with Anil as my advisor over the course of my doctoral studies. His insights and guidance helped me in my journey of becoming an independent researcher. I would also like to thank Phil who's advice was invaluable in both professional and academic matters. In addition I would like to thank the other members of my dissertation and qualifying committee Peter Bartlett and Alper Atamturk for their time and service.

I could have not completed my PhD without the support of my loving parents Liat and Eli Mintz. Their perspective and advice was extremely helpful and gave me the motivation I needed even through the rough patches. I would also like to thank my brothers Idan and Michael Mintz for their support through the process.

Over the course of my PhD I have had the opportunity to collaborate with many individuals outside my committee, including Yoshimi Fukuoka who's experience and advice were influential in helping me begin to understand the complexities of healthcare research. In addition to Yoshimi, I am also grateful to Elena Flowers and the rest of the team at the UCSF school of nursing for all the support and collaboration through the years. I would also like to thank Joanna R. Pedrasa and the rest of the team from the University of the Phillipines and the PCARI project for the opportunity to learn from their diverse perspectives and work on fascinating research.

Through the process of obtaining my PhD I have made many great friends who have helped influence and challenge my identity as a scholar. For the many discussions and collaborations I would like to thank my lab mates Auyon Siddiq, Mo Zhou, Matt Olfat, and Pedro Hespanhol. I would also like to thank the many other fascinating and wonderful students who I've had the pleasure of meeting in the IEOR department including: Chen Chen, Andrés Goméz , Wei Qi, Dean Grosbard, Carlos Deck, Alfonso Lobos, Salar Fatahi, Georgios Patsakis, Mark Valendintski, and Erik Berteli. In addition I would like to thank the staff of the IEOR department including Keith and Anayancy for their support and organization. Finally, I would like to thank professors Zuo-Jun Max Shen, Dorit Hochabaum, Candace Yano, and Shmuel Oren for their advice and helpful conversations.

Besides my involvement in the department, I have had the pleasure of working with and meeting many great individuals while living in the Bay Area. I would like to thank my colleagues from the graduate assembly for all of our late nights and advocacy including: Quico Spaen, Andy Schwartz, Thomas Gilbert, Arman Jabbari, and David Fridovich-Kiel. I would also like to thank my friends and former coworkers for their social support and many adventures including: Andrew, Jeffery, Peter, Savanah, Beth, Austin, David, Michael, Greg, Courtney, Joshua , Mark, and Anna.

# Chapter 1

# Introduction

Over the past few years, there has been a significant increase the availability of data and computing resources that has significantly changed how individuals interact with analytics technology. For instance, services such as next word text prediction, fitness activity tracking, automated climate control, and targeted advertising have become almost completely integrated into daily life. In general, each of these services requires two essential components: a predictive model trained using data and a recommendation algorithm which uses this model to perform the desired function. One of the key contributing factors for the success of these technologies is the increased adoption of smart devices such as smart phones, wearable devices, and other smart sensors. These devices allow for cheep and simple ways to both deploy recommendation algorithms and collect the data needed to train the predictive models.

In particular, these new data sources and implementation strategies have created what we will refer to as precision service systems. In general, a precision service system can be thought of as any system which constitutes a large group of heterogeneous agents where the overall performance of the system depends on how effectively policies are set for each individual agent. These systems can be described abstractly as a the scenario of a single decision maker (e.g. the ad server, fitness app etc.) attempting to influence a group of agents (i.e. the users) towards a particular desired goal by computing a policy or intervention. This decision maker can in general be thought of as having only partial information of how these agents will react and may have a limited amount of resources at their disposal to implement the intervention. Several real world applications that can be thought of as precision service systems include personalized healthcare, targeted advertising, demand response contracting, among other examples.

As these precision systems become more common place however, new statistical estimation and optimization tools need to be developed to account for how they effect and interact with users. For instance, in the realm of online advertising, initially one of the most successful developments was the use of individual browsing data to target advertising displays. This however meant that users were exposed to a slew of similar ads and eventually stopped paying attention to the advertising campaigns in an effect called banner blindness. Similarly, in the realm of fitness tracking, currently most devices and applications provide exercise goals to users in order to motivate them to keep performing physical activity. In general, the goals are computed using either the entire user population's metrics or through an interpretation of government provided guidelines, and generally do not vary from day to day. This may result in exercise goals that are either too easy or too difficult to accomplish for different users resulting in decreased interest and participation.

These examples, among others, indicate that to design better tools for optimizing precision service systems the following engineering challenges need to be addressed:

1. The models used to mathematically describe individual behavior need to be adjusted to effectively capture the nuances of decision making.

2. Since recommendations are made at the individual levels, these models must be compatible with individual level data streams, such as those provided by smart devices, that will be effected by the recommendations provided.

3. The way the parameters of these models are estimated and recommendations calculated needs to be scalable to a large population size.

The goal of this thesis is to begin to address these challenges by developing a mathematical framework of precision analytics. We will first address the problem of modeling agent behavior by using concepts from the social sciences and behavioral theory to augment existing adaptive prediction models. In particular, we examine how modeling agents using utility theory can lead to effective models that can capture the complex distributions that arise in these behavioral settings. Then we examine how these models can be used for optimization. In particular we show that these approaches have both useful theoretical properties as well as good performance in an application setting.

The remainder of the thesis will proceed as follows. In Chapter 2 we describe how behavioral models can be developed, using personalized healthcare as the principal application. In this chapter we characterize a behavioral model that can be used to both estimate model parameters and impute values of missing data points. We expand this model to show that its predictive power can increase using prior information in a non-parameteric fully Bayesian framework. We further show that these models can be trained fairly quickly using a mixed integer linear program (MILP) and standard commercial solvers. To validate this approach, we compare the performance of this model against standard machine learning methods and show that this modeling methodology is competitive with these state of the art methods. The contents of this chapter roughly correspond to the material in our paper (Aswani et al., 2016).

In Chapter 3 we generalize the behavioral model and consider a precision analytics setting with infrequent and costly decisions. In this setting, the decision maker has a budget on how they can implement their policy, and only require recomputing their policy on a weekly or monthly basis. We first perform a statistical analysis of the behavioral model from the previous chapter and show that it has desirable properties including statistical consistency. We then develop a two stage algorithm to compute an asymptomatically optimal policy for this particular problem setting. This approach is then evaluated using a simulation study showing how it can be used to schedule clinical appointments and set exercise goals in the context of clinically supervised weight loss programs. The simulation results show that using this type of precision analytics approach is more effective then simple heuristics that could be used for policy design. The contents of this chapter correspond to the material in our paper (Mintz et al., 2017a).

In Chapter 4 we consider a different precision analytics setting which requires frequent but relatively cheap decisions. In this setting the decision maker may not have a cost or budget on how to implement their policy but must make a policy decision on a daily or more frequent scale. We first discuss how this setting can be thought of as a particular case of a non-stationary restless bandit problem which we call the reducing or gaining unknown efficacy (ROGUE) bandit problem. We then develop an upper confidence bounds approach to approximately solve this bandit problem and show that it obtains an efficient rate of expected regret. We then perform computational experiments and show how this approach can be used effectively in two different problem settings. One of our experiments shows that using this methodology could be effective in a precision healthcare setting. The materials in this chapter correspond to the contents of our paper (Mintz et al., 2017b).

# Chapter 2

# Modeling Agent Behavior

## 2.1 Introduction

Effective design of systems involving human agents often requires models that characterize the agents' varied responses to changes in the system's states and inputs. Most operations research (OR) models quantify agent behavior as decisions generated by optimizing static utility functions that depend upon time-varying system states and inputs. In contrast, researchers in the social sciences have found that the motivational psychology of agents changes in response to past states, decisions, and inputs from external agents (Kanfer, 1975, Ajzen and Fishbein, 1980, Gonzalez et al., 1990, Janz and Becker, 1984, Joos and Hickam, 1990, Bandura, 2001); however, these social science models are primarily qualitative in nature, making them challenging to incorporate into OR design and optimization approaches. In this chapter, we focus on developing a predictive modeling framework that incorporates time-varying motivational states (which describe the changing efficiency or preferences of the agent) – thereby quantifying agent behavior as decisions generated by optimizing utility functions that depend upon time-varying system states, system inputs, and motivational states, all evolving according to some modeled process based on qualitative social science models of behavior change.

Our ultimate goal is to solve optimization problems to more effectively allocate resources in systems with human agents; to do this we need to develop behavioral models that can be integrated as constraints in standard optimization approaches. In this chapter, we develop a modeling framework that inputs noisy and partially-missing data and uses this to estimate the parameters of a predictive model consisting of (a) a utility-function describing the decision-making process that depends upon time-varying system states, system inputs, and motivational states, and (b) temporal dynamics on agent's system state and motivational state (i.e., often referred to as the type of the agent). We consider two distinct but related kinds of estimates: estimation of the set of parameters for the utility function and dynamics, and separately, estimation of the distribution of future states.

The framework we develop in this chapter is described within the context of modeling the behavior of individuals in a weight loss program; specifically, we are interested in using a short time-span (e.g., 15-30 days) of physical activity and weight data from an individual participating in a weight loss program in order to effectively characterize the likelihood of whether or not that individual will achieve clinically significant weight loss (i.e., 5% reduction in body weight) after a long period of time (e.g., 5 months). While machine learning approaches such as support vector machines (SVMs) (Hastie et al., 2009, Wang et al., 2017, Oztekin et al., 2018) and artificial neural networks can be used to make binary predictions of significant weight loss based on a short time span of data (Hastie et al., 2009) they have two significant limitations: first there is no obvious way to integrate them into an optimization model, and second these approaches are generally limited in their interpretability (Breiman et al., 2001). Here, we show that in contrast to these machine learning methods, our approach is interpretable since the equations are based on models from the social sciences, and can be incorporated into optimization models since it is posed as a mixed integer linear program (MILP), while maintaining comparable prediction accuracy.

### 2.1.1 Personalized Treatments and Obesity

Obesity is a significant problem in the United States. About 70% of American adults are overweight or obese (Flegal et al., 2012), and its annual cost to the health care system is estimated to be $350 billion (Valero-Elizondo et al., 2016). Currently, the most effective treatments for obesity are weight loss interventions composed of counseling sessions by clinicians and daily goals for physical activity and caloric consumption. The Diabetes Prevention Program Research Group (2002, 2009) showed that participating in these types of treatments results in significant weight loss of 5-7% and can prevent the onset of type-2 diabetes with few side effects. However, adherence to these clinician-set goals decreases over time (Acharya et al., 2009), and these programs are labor-intensive and expensive to sustain (McDonald et al., 2002, Diabetes Prevention Program Research Group, 2003). Making these interventions more *effective and efficient* will require designing treatments personalized to each individual's preferences.

While individualized goal-setting and personalized interventions are crucial to the success of these programs, these features are expensive to provide. Cost efficient programs will need automation of goal-setting and scheduling of counseling resources for individuals to succeed in reducing their weight. Such approaches will likely involve digital/mobile/wireless technologies, which already have high adoption rates (Lopez et al., 2013, Bender et al., 2014) and have shown promise for improving the quality of and adherence to weight loss programs (Fukuoka et al., 2011). These technologies allow clinicians and researchers to remotely collect real-time health data and communicate with individuals participating in the program. However, healthcare data sets generated by mobile devices have been underutilized to date, and little research has focused on effective ways to utilize individuals' health-related data patterns to improve and personalize weight loss interventions (Fukuoka et al., 2011, O'Reilly and Spruijt-Metz, 2013, Pagoto et al., 2013, Azar et al., 2013).

### 2.1.2 Overview

Ultimately, effective automated approaches will depend upon nuanced models to predict the effects different interventions (i.e., changes in activity and caloric goals, or specific types of counseling) will have on the weight loss trajectories of different individuals. In this chapter, we present an initial step – specifically, we develop an approach for using a short time-span (e.g., 15-30 days) of physical activity and weight data from an individual participating in a weight loss program to effectively characterize the likelihood of whether or not that individual will achieve clinically significant (i.e., 5% reduction in body weight) weight loss after a long period of time (e.g., 5 months) as a function of the physical activity goals and amount of counseling given to the individual. (The Diabetes Prevention Program Research Group (2002, 2009) showed 5% weight loss provides substantial health benefits.) As discussed above, this type of predictive tool will ultimately enable the adaptive design of more effective and cost efficient interventions. Towards this end, we also show how our predictive model is able to predict the impact of changes in the intervention treatment on the weight loss trajectory of a specific individual.

A key feature of predicting future behavior is the inherent uncertainty due to having limited data. As a result, it is natural to consider predictive modeling approaches that generate ranges or intervals of predictions. Though frequentist approaches can be used to construct confidence intervals, we instead propose a Bayesian approach that constructs a range of predictions characterized by a *posterior* distribution. An important benefit of our Bayesian (as compared to a frequentist) approach is that it can incorporate data from individuals that have been in the program for a longer period of time or have even completed a fixed duration (e.g., 5 months) of the program. We quantitatively show in Section 2.6 that incorporating the information of other individuals using a nonparametric Bayesian prior distribution improves the accuracy of predictions versus not using a Bayesian framework.

Our resulting predictive modeling approach is presented in Section 2.5. In the preceding sections, we develop essential elements for constructing the model. We first describe the structure of mobile phone-based weight loss interventions in Section 2.2. Section 2.3 describes our utility-maximizing model of the decisions of an individual participating in a weight loss intervention. Mathematically, we represent prior information in the Bayesian framework as histograms of parameter values for the utility functions of individuals that have completed the fixed duration of the program. To compute these parameters, we solve a maximum likelihood estimation (MLE) problem, which is the focus of Section 2.4. Our predictive modeling approach in Section 2.5 uses the utility-maximizing framework and corresponding histograms of parameter values to predict the weight loss trajectory of a single individual. Both the MLE in Section 2.4 and predictive model in Section 2.5 are computed by solving a mixed integer linear program (MILP).

To validate our predictive modeling approach, we use a longitudinal data set collected from a 5-month randomized controlled trial (RCT) of a mobile phone-based weight loss program. Section 2.6 begins with an overview of this RCT, and additional details are available in Fukuoka et al. (2015). Next, we evaluate the effectiveness of our approach for predicting whether or not an individual will achieve clinically significant (i.e., 5% or more) weight loss at the end of the intervention. We validate our approach by showing its binary predication accuracy is comparable to standard machine learning methods (i.e., linear SVM, decision tree, and logistic regression) in terms of prediction quality. In contrast to these machine learning methods, our predictive model is also able to determine the impact of changing intervention parameters for a specific individual on that individual's weight loss trajectory, and we conclude with a discussion of this aspect of our model and how it can be used to perform optimization.

### 2.1.3 Literature Review

Statistical classification methods (which include logistic regression, support vector machines, neural networks, and random forests) predict a binary $\{-1, +1\}$ output label based on an input vector (Hastie et al., 2009, Denoyel et al., 2017). In the context of weight loss interventions, these approaches could predict whether $(+1)$ or not $(-1)$ an individual will achieve 5% weight loss after 5 months, based on 30 days of an individual's data. However, these approaches lack interpretability (Breiman et al., 2001) and cannot be incorporated as constraints into standard optimization approaches. Our predictive modeling approach is similar in that it can be used as a classifier (i.e., it can predict whether or not an individual achieves 5% weight loss), but it differs in that its equations are based on models from the social sciences, and can be incorporated into optimization models since it can be posed as a mixed integer linear program (MILP), making it more applicable for addressing the problem of intervention design.

A number of predictive models have been developed to determine the impact of changing a medical intervention on the health outcome for an individual, including: Markov chain models (Ayer et al., 2012, Mason et al., 2013, Deo et al., 2013, Andersen et al., 2017), dynamical systems models (Helm et al., 2015), decision tree models (Wu et al., 2013), graph-theoretic models (Fetta et al., 2018), bandit models (Negoescu et al., 2014), and dynamic programming models (Engineer et al., 2009). (This literature also studies the problem of designing optimal treatment plans, which we do not consider in our present paper.) Our work is similar in that we develop an approach to predict future body weight of an individual as physical activity goals and counseling scheduling are changed. One key difference is in the data available in weight loss programs. Existing approaches are designed for situations where data is collected infrequently (e.g., only during clinical visits), whereas in weight loss programs the data is collected daily using mobile devices. Our work seeks to develop a predictive modeling approach that can leverage this increased data availability in order to make improved predictions. Moreover, existing approaches focus either on motivational states (Mason et al., 2013) or health states (Ayer et al., 2012, Deo et al., 2013, Helm et al., 2015, Wu et al., 2013, Negoescu et al., 2014, Engineer et al., 2009). We seek to combine the notions of motivational and health states into a single predictive model, which is a modeling approach that has not been previously considered.

Previous approaches for automated exercise and diet management significantly differ in the goal of the predictive modeling. Bertsimas and O'Hair (2013) develop a system that learns a predictive model of an individual's dietary preferences and then designs a plan of what food to eat and how much time to exercise to maintain low blood glucose levels. The output of this predictive model is blood glucose levels and satisfaction of a given dietary plan, whereas we are interested in making predictions regarding future body weight. Additionally, this predictive model does not consider adherence to the prescribed plans (e.g., the individual may overeat or may not exercise the amount indicated by the plan), whereas our approach quantifies the level of adherence to prescribed physical activity goals and guidance on caloric intake. The Steptacular program (Gomes et al., 2012) used monetary incentives to encourage individuals to walk more, but a predictive model was not developed to design the incentives; our approach differs in that we seek to build a predictive model so that in the future we may be able to optimize the weight loss intervention for each individual.

### 2.1.4 Contributions

We develop a number of novel optimization modeling and analysis techniques that we believe will be useful for expanding the scope of predictive models of human decision-making in complex systems. For instance, much mobile phone data contains non-negligible noise and suffers from missing data points (Chen et al., 2012). Aswani et al. (2018) showed that statistically consistent estimation of model parameters in a utility-maximization framework requires joint estimation of the missing data and model parameters. It is known (see for instance Bickel and Doksum (2006)) that such joint estimation does not represent statistical over-fitting, and in fact all regression approaches (even basic linear regression) jointly provide estimates of denoised data and model parameters; however, only the model parameter estimates are statistically consistent (Bickel and Doksum, 2006, Aswani et al., 2018). Existing approaches for dealing with missing data (e.g., the EM algorithm (Hastie et al., 2009)) generate an estimate by computing the local optimum of a suitably defined optimization problem that computes the parameters of the predictive model. Instead, we construct optimization models formulated as mixed integer linear programs (MILP's) that are able to simultaneously estimate missing/noisy data and parameters of the utility-maximizing framework; this yields global optima of the parameter computation optimization problem.

As mentioned above, we can likely improve trajectory predictions for a specific individual in a weight loss intervention by leveraging mobile phone data from other individuals who have already completed the intervention. This challenge can be posed in a Bayesian framework, but existing nonparametric approaches require computing numerically challenging integrals. In this chapter, we provide what is to the best of our knowledge the first Bayesian estimation approach in which the prior distribution is purely data-driven and described by a histogram. For this Bayesian estimation, we use integer programming, and we show that a data-driven distribution can be represented as a piecewise constant function, which can then be formulated within a MILP (Vielma, 2015).

In many cases, patients favor behavior that does not improve (or is not optimal with respect to) their health outcomes. Non-adherence to a medical plan falls in to this category. Social scientists sometimes label such behavior "irrational" (Brock and Wartman, 1990); however, an argument has been made that many instances of "irrational" behavior are in fact rational decisions when considering a patient's actual utility function (Gafni, 1990, Cawley, 2004). In our case, we explicitly use a utility function in which the individual is assumed to heavily discount future health states, a behavior that is often characterized as "irrational" (Brock and Wartman, 1990). We note, however, that while these modeling choices may be controversial, the particular utility function framework that we develop has an alternative interpretation that does not make reference to utility maximization. In particular, our approach can alternatively be interpreted as leading to a model that has the best theoretical predictive accuracy given the set of underlying equations that characterize this framework. For additional details on this interpretation please see (Aswani et al., 2018). Thus, even if the behavioral argument we advance in subsequent sections of this chapter does not accurately capture individuals' behavior, the framework we describe still enables us to make the most accurate set of predictions possible using the set of equations underlying the predictive model.

## 2.2 Structure of Mobile Phone-Based Weight Loss Interventions

Currently the healthcare community is refining a new class of weight loss interventions that rely on mobile phones and digital accelerometers (Gomes et al., 2012, Fukuoka et al., 2015, Flores Mateo et al., 2015). Though the particular features of these programs often differ, there is a growing consensus on the broad structure of these programs. In general, each individual is provided with (i) a mobile phone app and a digital accelerometer, and (ii) in-person counseling sessions. The digital accelerometer is used to measure daily physical activity, and the digital aspect of the device simplifies data sharing and data uploading. The mobile phone app delivers physical activity goals, educational messages (such as those from (Diabetes Prevention Program Research Group, 2002, 2009)), and provide an interface for individuals to enter dietary and body weight information.

The accelerometer measures the number of steps taken each day since the majority of exercise for individuals in such weight loss interventions consists of walking. Individuals are also typically asked to input weight measurements multiple times a week into the mobile app. In principle, the data available for each individual consists of daily weight and step amounts; however, data for some dates is missing because individuals forget to enter weight data into the mobile app, wear the accelerometer, or because of a technical problem with the app. The age, gender, and height of each individual is also known data in these programs.

Individuals participating in such mobile phone-based weight loss interventions receive additional interaction. After an initial baseline period, exercise goals in terms of a minimum daily step count are provided to each individual. The goals change at regular intervals (e.g., every week). Individuals also have office visits (or phone calls) at regular intervals, during which they received behavioral counseling about their nutritional choices and physical activity. The exercise goals and timing of the office visits (or phone calls) are set in advance, and thus are also known data in these programs.

## 2.3 Formulating the Utility-Maximizing Framework

The utility-maximizing framework we propose has two components. The first describes how an individual makes decisions regarding the amount of steps and caloric intake, and this is formulated in terms of a utility-maximizing individual. The utility function contains heavy discounting of future health states, a behavior that is often characterized as “irrational” (Brock and Wartman, 1990). The second describes how the individual’s weight and *type* (a set of parameters describing each individual) evolve over time as a function of current states and decisions. This second part is formulated in terms of a linear dynamical system.

### 2.3.1 Summary of Framework

A subscript $t$ denotes the value of a variable on the $t$-th day. Let $f_t \in \mathbb{R}_+$ denote the amount of calories consumed, $u_t \in \mathbb{R}_+$ be the number of steps, $w_t \in \mathbb{R}_+$ be the weight of the individual, $g_t \in \mathbb{R}_+$ be the given exercise goal in terms of number of steps, and $d_t \in \{0, 1\}$ indicate whether or not an office visit occurred. We refer to $\theta_t = (k, q, s_0, s_t, p_t, \mu)$ as the *type* of the individual. The parameters $a, b, c, k \in \mathbb{R}$ describe the weight dynamics, are based on the physiology of the individual, and can be precomputed based on the age, gender, and height of the individual (Mifflin et al., 1990). Another set of the parameters are used in the utility function. These include $r_f, r_u \in \mathbb{R}$ which represent the marginal utility of quadratic terms, $q, s_0$ which represent baseline preferences in terms of physical activity and caloric consumption respectively, $p_t \in \mathbb{R}$ which represent the marginal dissutility of failing exercise goals, and $s_t \in \mathbb{R}$ which represents the current preference of caloric consumption. The last set of parameters describe the type dynamics, including $\mu \in \mathbb{R}_+$ that captures the impact of achieving an exercise goal, and $0 < \gamma < 1$ which is a discount factor representing the diminishing effect of the intervention over time. The $\beta_t, \delta_t \in \mathbb{R}_+$ are random variables with finite variance that represent the impact of an office visit, and $z_t \in \mathbb{R}$ is a zero-mean random variable with finite variance that denotes weight fluctuations from unmodeled effects. These random variables $\beta_t, \delta_t, z_t$ are individual-specific, but we do not consider them to characterize the *type* of the individual. This is because we assume their distributions are the same for each individual, and so the expected behavior of any particular individual will not depend in a unique way upon these random variables. Using these quantities, we define the following utility functions and dynamics.

1. Individual decision-making when no exercise goals are given is

$$\begin{aligned}(u_t, f_t) = \arg\max_{u,f} \; & -w_{t+1}^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t \\ \text{s.t. } \; & w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k.\end{aligned} \qquad \mathbf{U}_{\textbf{no goals}}$$

   Individual decision-making when exercise goals are given is

$$\begin{aligned}(u_t, f_t) = \arg\max_{u,f} \; & -w_{t+1}^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t + p_t \cdot (u_t - g_t)^- \\ \text{s.t. } \; & w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k.\end{aligned} \qquad \mathbf{U}_{\textbf{goals}}$$

   Note that $\mathbf{U}_{\textbf{no goals}}$ and $\mathbf{U}_{\textbf{goals}}$ refer to the $(u_t, f_t)$ that are computed by solving the corresponding optimization problems.

2. Weight and type are assumed to evolve according to the following:

$$w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k + z_t \tag{2.1}$$

$$s_{t+1} = \gamma \cdot (s_t - s_0) + s_0 - \beta_{t+1} \cdot d_{t+1} \tag{2.2}$$

$$p_{t+1} = \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu \cdot \mathbb{1}(u_t \geq g_t). \tag{2.3}$$

   Observe that the time index in (2.2), (2.3) for $\beta, \delta, d$ is $t+1$ because we assume that the impact of a clinical visit occurs on the day of the visit.

Note that in $\mathbf{U_{no\ goals}}$ the caloric consumption preference $s_t$ is time-varying, whereas the physical activity preference $q$ is constant. The reason is that in clinically-supervised weight loss programs, individuals are encouraged to reduce their caloric consumption at the beginning of the program – in contrast, the individuals are asked to not increase their physical activity level until they begin to receive goals (Fukuoka et al., 2011). Thus, our predictive model assumes that the physical activity preference remains constant during the period in which no goals are given.

### 2.3.2 Structure of Utility Function

We assume an individual's utility function is separable with respect to weight, caloric intake, and exercise amount. An individual with *perfect knowledge* of his or her type $\theta_t$ may choose their exercise amount $u$ and caloric intake $f$ to maximize a utility of the form $\sum_{k=0}^{\infty} \alpha^{-k} \cdot \mathbb{E}(U_1(w_{t+k+1}, d_{t+k}, g_{t+k}; \theta_{t+k}) + U_2(u_{t+k}, d_{t+k}, g_{t+k}; \theta_{t+k}) + U_3(f_{t+k}, d_{t+k}, g_{t+k}; \theta_{t+k}))$, subject to weight $w_{t+k+1} = \eta(w_{t+k}, u_{t+k}, f_{t+k}, \xi_{t+k})$ and type dynamics $\theta_{t+k+1} = \zeta(\theta_{t+k}, w_{t+k}, u_{t+k}, f_{t+k}, \xi_{t+k}, d_{t+k}, g_{t+k})$, where $\xi_{t+k} = (z_{t+k}, \beta_{t+k}, \delta_{t+k})$ are random variables, $\alpha \in [0, 1)$ is a discount factor, $U_1, U_2, U_3$ are utility functions, and $\eta, \zeta$ are functions that define the dynamics. Note that utility depends on weight one day ahead of the corresponding decision because future weight and present decisions affect utility.

However, it is not true that individuals make health care decisions with the goal of maximizing long term health benefits. Indeed, it is common for individuals to very heavily discount the impact of present decisions on future health outcomes (Chapman and Elstein, 1995). To capture this behavior that is sometimes characterized as "irrational" (Brock and Wartman, 1990), we explicitly use a utility function in which the individual is assumed to heavily discount future health states.

**Proposition 2.1.** If the discount factor is $\alpha = 0$, then this is equivalent to an equation where the individual makes a decision considering only the one-day impact: $\max_{u,f}\ \{\mathbb{E}(U_1(w_{t+1}, d_t, g_t; \theta_t) + U_2(u_t, d_t, g_t; \theta_t) + U_3(f_t, d_t, g_t; \theta_t)) \mid w_{t+1} = \eta(x_t, u_t, f_t, \xi_t)\}$.

A complete proof can be found in Appendix A.1, however note that this result can be reached using direct computation.

### 2.3.3 Choice of Utility Function

Corresponding terms in the utility function are chosen to match to particular behaviors expected by social cognitive theory (Bandura, 2001): In this context, social cognitive theory asserts that caloric consumption and physical activity depend upon (1) *self-efficacy*, which is an individual's belief in their ability to achieve positive behavioral changes and is characterized by the coefficients $p_t, q, s_t$; and depend upon (2) receiving a positive reward from a small amount of weight loss for engaging in positive behavioral changes. We choose $U_1 = -w_{t+1}^2$, $U_3 = -r_f f_t^2 + s_t f_t$, $U_2 = -r_u u_t^2 + q u_t$ if no goal is given, and $U_2 = -r_u u_t^2 + q u_t + p_t \cdot (u_t - g_t)^-$ if a goal is given. Dislike for large amounts of steps and caloric intake is captured by the $-r_u u_t^2$ and $-r_f f_t^2$ terms. Positive satisfaction for increasing steps and caloric intake is represented by the $q u_t$ and $s_t f_t$ terms. An individual's preference for lower weight is reflected by the $-w_{t+1}^2$ term. And an increase in satisfaction for getting closer to the exercise goal is captured by the $p_t \cdot (u_t - g_t)^-$ term.

**Remark 2.1.** Observe that as $p_t$ increases, the utility of meeting a step goal increases, and as $s_t$ increases, the utility of higher caloric intake increases. Thus, we can interpret the values $p_t, s_t$ as a quantification of the adherence of an individual to step goals and dietary goals, respectively.

**Remark 2.2.** An alternative choice is $U_1 = -w_t^2 + 2w_b w_t$, which has an additional linear term with coefficient $w_b$. After completing the square, this is equivalent to choosing $U_1 = -(w_t - w_b)^2$, which makes its interpretation clear: The $w_b$ coefficient should be interpreted as the preferred weight of that individual. From a computational standpoint, $w_b$ can be estimated using the same approach that we describe in later sections for estimation of $r_f, r_u$. However, we chose to not include the linear term for two reasons. The first is that including this linear term does make estimation more slow computationally. The second is that choosing $w_b = 0$ for all individuals (which makes $U_1 = -w_t^2$) and then scaling for each individual the other coefficients in $U_2, U_3$ can reasonably approximate within a finite range of weights a $U_1$ with a linear term. Our second reason also explains why a purely linear $U_1 = -w_t$ is not an appropriate choice, because a purely linear $U_1$ cannot capture the diminishing returns to weight loss as weight decreases towards the desired weight. As we will show later, setting $w_b = 0$ leads to accurate predictions, which ultimately validates our choice.

While other functional forms can represent the behaviors expected by social cognitive theory, these choices have several advantages. The choice that positive utility ($q u_t$ and $s_t f_t$) increases at a slower rate than disutility decreases ($-r_u u_t^2$ and $-r_f f_t^2$) ensures that an individual takes a finite number of steps and consumes a finite amount of calories. (Other choices can lead to a situation where the individual is predicted to take an infinite number or steps or consume an infinite number of calories, which is clearly unreasonable.) Moreover, these choices ensure the objective is strictly concave, which ensures that an individual is predicted to make only one decision; if the utility function was merely concave, then there may be multiple maximizers that correspond to a set of different possible decisions on the number of steps and calories.

Additionally, this functional form has a relatively low parameter count, which facilitates estimation. For instance, there is no linear term for weight $w_t$. The utility term $qu_t$ is kept constant because explicitly incorporating an increase in exercise utility (with an office visit) would be an over-parametrization due to the $p_t \cdot (u_t - g_t)^-$ term. Furthermore, we do not need to include a parameter for the $-w_t^2$ term because this would simply scale the function, and would not change the decision. Lastly, our choice implies that goal setting has no impact beyond the goal amount.

**Remark 2.3.** Restated, the utility term $p_t \cdot (u_t - g_t)^-$ is at its maximum value for all $u_t \geq g_t$. This is a simplification to reduce the number of terms. A more detailed framework would also incorporate positive utility for exceeding the goal, such as by including the term $\rho_t \cdot (u_t - g_t)^+$. The reason we do not include a linear term $\rho_t \cdot (u_t - g_t) = \rho_t \cdot u_t - \rho_t \cdot g_t$ is that such a term inherently cannot capture the satisfaction of meeting a goal, because it has the same effect (due to $\rho_t \cdot g_t$ being a constant) as including the term $\rho_t \cdot u_t$.

### 2.3.4 Dynamics of Weight

We also need to specify weight dynamics. Standard physiological arguments (i.e., weight change is proportional to "calories-in minus calories-out") imply that the weight dynamics are given by $w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k + z_t$, where $a, b, c, k \in \mathbb{R}$ are coefficients that can be computed using existing physiological models, and $z_t$ is a zero-mean random variable that captures unmodeled changes in weight (e.g., water fluctuation, physical activity in addition to steps, etc.). Suppose $w_t, k, z_t$ are specified in units of kilograms, $f_t$ is specified in units of kilocalories (also known as dietary calories), and $u_t$ is specific in units of steps. Then a derivation given in the A.2 and based on the Mifflin St Jeor Equation (Mifflin et al., 1990) for the basal metabolic rate (BMR) gives $a = 0.9987$ and $k = -8.0357 \times 10^{-4} \cdot h + 6.4286 \times 10^{-4} \cdot a + s$, where $h$ is height in centimeters, $a$ is age in years, $s = -6.4286 \times 10^{-4}$ for males, and $s = 2.0700$ for females. To compute $b$, we note that 2000 steps is roughly equal to walking one mile and consumes about 100 calories, largely independent of the height, weight, age, and gender of an individual (Hill et al., 2003). This gives a value of $b = -6.4287 \times 10^{-6}$. Last, the value of $c = 1.2857 \times 10^{-4}$ is computed by performing the unit conversion that 3500 calories is 0.45 kilograms.

One consequence of linear weight dynamics is simplification of the utility-maximizing framework:

**Proposition 2.2.** When the weight dynamics are linear, as in (2.1), we can rewrite the objective of the utility-maximizing framework as $-(a \cdot w_t + b \cdot u_t + c \cdot f_t + k)^2 + U_2(u_t, d_t, g_t; \theta_t) + U_3(f_t, d_t, g_t; \theta_t) - \mathbb{E}(z_t^2)$.

A complete proof of this proposition can be found in Appendix A.1 but here we will provide some intuition. Observe that the only stochasticity in the utility function is in $z_t$, which has zero mean and cannot be used for decision making at time $t$. This means that the terms with $z_t$ in the objective function have zero expectation and can therefore be eliminated.

**Remark 2.4.** The main insight from this substitution is that decisions made by an individual following the utility-maximizing framework do not depend on the stochasticity because $-\mathbb{E}(z_t^2)$ is a constant that does not depend on the decisions.

Before describing the type dynamics, we discuss a more detailed model for the weight dynamics. Specifically, a phenomenon known as *adaptive thermogenesis* (Doucet et al., 2001, Rosenbaum et al., 2008) causes the metabolism of an individual who has lost weight to decrease. Our weight dynamics (2.1) can be modified to incorporate this phenomenon by allowing the $z_t$ to have a non-zero mean. Though we do not use this more detailed model in this chapter, we briefly outline how our MLE and Bayesian prediction formulations (that will be described in upcoming sections) would change: The first change is that the $k$ term in the constraints would be replaced with $k + m_t$, where $m_t$ is a new variable that represents the mean of $z_t$. This change allows the $z_t$ in our formulations to have a non-zero mean. The second change is that an additional constraint $\sum_{t=1}^{n-1} |m_{t+1} - m_t| \leq \sigma_m$ is added to our formulations, where $n$ is the time step at which we are solving the formulation and $\sigma_m$ is a constant the bounds the amount of metabolism change, and this constraint is known as *fused lasso* (Tibshirani et al., 2005) in the statistics and machine learning literature. This additional constraint has been show to have properties (Tibshirani et al., 2005) that would lead to estimates that ensure the estimated change in metabolism becomes roughly constant after an individual's weight stops changing, which is an important property because it matches what is clinically observed with changes in metabolism after weight loss (Doucet et al., 2001, Rosenbaum et al., 2008).

### 2.3.5 Dynamics of Type

The type dynamics are as specified in (2.2),(2.3), where $\gamma, s_0, \mu$ are scalars and $\beta_t, \delta_t$ are random variables. Specific terms in these dynamics correspond to principles of social cognitive theory, which says in this context that self-efficacy as quantified by $s_t, p_t$ will increase in response to social contact during office visits and in response to successfully achieving past goals. The uncertain impact of office visits is modeled by the stochastic $\beta_t$ and $\delta_t$. The fact that office visits sometimes make external goal-setting more effective and decrease interest in eating is described by the $\delta_{t+1} \cdot d_{t+1}$ and $-\beta_{t+1} \cdot d_{t+1}$ terms, respectively. Because the impact of a single office visit decreases to zero over time, the dynamics include the terms $\gamma \cdot (s_t - s_0) + s_0$ and $\gamma \cdot p_t$. Observe that these discounting terms are different because $s_0, q$ are the baseline preferences for caloric consumption and physical activity, respectively. So the first discounting term ensures $s_t$ goes to $s_0$ without more office visits, and the second discounting term ensures $p_t$ goes to zero without more office visits since $q$ already encodes the baseline coefficient for physical activity. Moreover, goal-setting can become more effective whenever the goal is met; this is characterized by the $\mu \cdot \mathbb{1}(u_t \geq g_t)$ term.

Multiple equation choices would lead to the behaviors suggested by social cognitive theory, but this set of choices ensures the dynamics are linear in $s_t, p_t$ and reduces the parameter count. The latter objective is achieved through (i) using the same parameter $\gamma$ for both the $\gamma \cdot (s_t - s_0)$ and $\gamma \cdot p_t$ terms, and (ii) using a constant parameter $\mu$ instead of allowing this to be a time varying quantity. Linearity in $s_t, p_t$ is important for favorable computational properties. Though the term $\mu \cdot \mathbb{1}(u_t = g_t)$ is nonlinear, it has special structure that allows efficient computation.

## 2.4 Maximum Likelihood Estimation (MLE) for Utility-Maximization

Estimating parameters of the utility-maximizing framework for a specific individual requires solving an optimization problem. However, formulating this model is challenging because the measurements suffer from noise and missing weight and step data. This can be overcome by formulating the optimization model so that its minimizer simultaneously estimates the values of weight, caloric intake, steps, type, and the random variables in the model for each individual. The optimization model for simultaneous estimation is generally a nonconvex, nonlinear program; and it is typical to generate an estimate by computing a local optimum (e.g., the EM algorithm (Hastie et al., 2009)). However, we show that simultaneous estimation can be modeled using as a MILP, allowing us to compute the global optimum of the optimization model for estimation.

We pose the estimation problem in the framework of MLE. Suppose that the data for a single individual consists of $(t_i, \tilde{w}_{t_i})$, for $i = 1, \ldots, n_w$, and $(\tau_i, \tilde{u}_{\tau_i})$, for $i = 1, \ldots, n_u$, where $n_w$ are the number of weight measurements, $n_u$ are the number of step measurements, and the noise model is $\tilde{w}_{t_i} = w_{t_i} + \nu_{t_i}$ and $\tilde{u}_{\tau_i} = u_{\tau_i} + \omega_{\tau_i}$, where $\nu_{t_i}, \omega_{\tau_i}$ are zero-mean random variables with finite variance. Note that the times $t_i, \tau_i$

do not coincide in general. Let $\psi_\nu(\cdot), \psi_\omega(\cdot), \psi_z(\cdot)$ by the probability density function (pdf) for the random variables $\nu_t, \psi_t, z_t$. The MLE problem seeks to estimate the type $\theta_t$ of each individual, using the above described data. It is important to further discuss the interpretation of the type $\theta_t$ that is estimated. Clearly there will be additional factors beyond the ones we have included in our predictive model that influence how an individual decides their daily caloric intake and number of steps, and so the measured data cannot be expected to exactly match our predictive model. In this context, the type $\theta_t$ that is estimated for each individual should be interpreted as those that maximize the prediction accuracy of the predictive model (Aswani et al., 2018) – a concept sometimes known as *risk consistency* in the statistics literature.

### 2.4.1 Initial Optimization Model for Computing MLE

Let $n = \max\{t_{n_w}, \tau_{n_u}\}$ be the number of days of data used for estimation, and let $m$ by the number of initial days before an exercise goal was given to the individual. For the utility-maximizing framework, the MLE is the minimizer of an optimization problem defined as

$$\begin{aligned} &\min\ \textstyle\sum_{i=1}^{n_w} -\log\psi_\nu(\tilde{w}_{t_i} - w_{t_i}) + \sum_{i=1}^{n_u} -\log\psi_\omega(\tilde{u}_{\tau_i} - u_{\tau_i}) + \sum_{t=1}^{n} -\log\psi_z(z_t) \\ &\text{s.t. } \mathbf{U}_{\textbf{no goals}}, (2.1) \text{ for } t = 1, \ldots, m-1; \quad \mathbf{U}_{\textbf{goals}}, (2.1), (2.2), (2.3) \text{ for } t = m, \ldots, n. \end{aligned} \qquad \mathbf{P}_{\textbf{mle}}$$

Recall that $\mathbf{U}_{\textbf{no goals}}$ captures decision-making without goals, $\mathbf{U}_{\textbf{goals}}$ captures decision-making model with goals, equations (2.1) are dynamics on weight, (2.2) and (2.3) are the dynamics of parameters $s_t, p_t$ respectively. Note that the first office visit is on the same day the first exercise goal is given. Since $s_t, p_t$ cannot change until the start of the intervention their dynamics begin at time $m$.

The problem $\mathbf{P}_{\textbf{mle}}$ is more challenging to solve than may initially appear. The variables $u_t, f_t$ are defined as the minimizing arguments of $\mathbf{U}_{\textbf{no goals}}$ and $\mathbf{U}_{\textbf{goals}}$. This makes the MLE the solution to a bilevel optimization problem (Dempe, 2002). Among the bilevel optimization problems that have been considered in the literature include inverse optimization with linear objectives (Ahuja and Orlin, 2001) and inverse optimization for combinatorial problems like assignment and spanning tree problems (Heuberger, 2004). In the context of bilevel optimization problems for estimating utility functions, approaches have been derived under the assumption of small noise (Keshavarz et al., 2011, Bertsimas et al., 2014); more recently, statistically consistent approaches for noisy measurements have also been proposed (Aswani et al., 2018). Here, we develop a new integer programming approach for solving our specific bilevel optimization problem in $\mathbf{P}_{\textbf{mle}}$.

### 2.4.2 Choosing the Distribution of Random Variables Representing Noise

We first must select the distribution of random variables representing noise $\nu_t, \psi_t, z_t$. Their variances $\sigma_1, \sigma_2, \sigma_3$ are constants that can be chosen based on our prior knowledge regarding the measurement accuracy of weight scales, measurement accuracy of accelerometers for measuring steps, and physiological information about the modeling errors of the Mifflin St Jeor Equation (Mifflin et al., 1990) for BMR. In our modeling, we used $\sigma_1 = 2$, $\sigma_2 = 0.1$, and $\sigma_3 = 0.1$.

Choosing zero-mean Gaussian random variables yields a quadratic objective for $\mathbf{P_{mle}}$: $\kappa_1 + \frac{1}{\sigma_1}\sum_{i=1}^{n_w}(\tilde{w}_{t_i} - w_{t_i})^2 + \frac{1}{\sigma_2}\sum_{i=1}^{n_u}(\tilde{u}_{\tau_i} - u_{\tau_i})^2 + \frac{1}{\sigma_3}\sum_{t=1}^{n}(z_t)^2$, where $\kappa_1$ is a constant. Alternatively, one could select $\nu_t, \psi_t, z_t$ to be zero-mean Laplace random variables, which have a pdf of $\psi(x) = \frac{1}{\sqrt{2\sigma}}\exp(-|x|/\sqrt{\sigma/2})$ with variance $\sigma$. The resulting objective of $\mathbf{P_{mle}}$ is proportional to $\sigma_1^{-1/2}\sum_{i=1}^{n_w}|\tilde{w}_{t_i} - w_{t_i}| + \sigma_2^{-1/2}\sum_{i=1}^{n_u}|\tilde{u}_{\tau_i} - u_{\tau_i}| + \sigma_3^{-1/2}\sum_{t=1}^{n}|z_t|$.

**Remark 2.5.** This resulting objective function becomes a linear objective function after a minor reformulation (see, for example, Section 6.1.1 of (Boyd and Vandenberghe, 2004)).

We assume the noise is Laplacian because this results in MILP optimization problems for estimation and prediction. Note that if we had assumed Gaussian noise, then this would have resulted in MIQP optimization problems for estimation and prediction. We have found that these resulting MIQP's are solvable using standard software, but that the prediction accuracy was not better than that of the MILP formulations arising from the Laplacian assumption. Hence we chose to assume Laplace noise because of the faster computation time for the resulting MILP's. The similar predictive accuracy under both assumptions is not surprising given that the difference in the objective is simply an absolute value of deviation versus the square of deviation.

### 2.4.3 Reformulating the MLE Using KKT

One approach to solving bilevel programs is to replace the convex optimization problems that are constraints by their corresponding necessary and sufficient optimality conditions (Dempe, 2002).

**Proposition 2.3.** Necessary and sufficient optimality conditions for $\mathbf{U_{no\ goals}}$ can be written as

$$\begin{aligned} &2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q = 0 \\ &2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_0 = 0. \end{aligned} \tag{2.4}$$

The complete proof of this proposition can be found in Appendix A.1 but here we will provide some intuition for this proposition. Essentially, the constraints of $\mathbf{U_{no\ goals}}$ can be eliminated by direct substitution, and equations (2.4) can be derived by computing the KKT conditions of the resulting optimization problem.

**Proposition 2.4.** Neccesary and sufficient optimality conditions for $\mathbf{U_{goals}}$ can be written as

$$\begin{aligned}
&2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q - \lambda_t^2 = 0 \\
&2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_t = 0 \\
&g_t - \epsilon - (g_t - \epsilon) \cdot x_t^1 \leq u_t \leq M + (g_t - \epsilon - M) \cdot x_t^1 \\
&(g_t - \epsilon) \cdot x_t^2 \leq u_t \leq M + (g_t + \epsilon - M) \cdot x_t^2 \\
&(g_t + \epsilon) \cdot x_t^3 \leq u_t \leq g_t + \epsilon + (M - g_t - \epsilon) \cdot x_t^3 \\
&0 \leq \lambda_t^2 \leq p_t; \qquad p_t - M \cdot (1 - x_t^1) \leq \lambda_t^2 \leq M \cdot (1 - x_t^3) \\
&x_t^1 + x_t^2 + x_t^3 = 1; \qquad x_t^1, x_t^2, x_t^3 \in \{0, 1\}.
\end{aligned} \tag{2.5}$$

The full proof of this proposition can be found in Appendix A.1 but here we will provide some intuition for the proof. To compute the optimality conditions of $\mathbf{U_{goals}}$, we can first reformulate problem as a quadratic program (QP). The resulting QP has a strictly concave objective and will satisfy constraint qualification meaning that the KKT conditions are both necessary and sufficient for optimality. After algebraic manipulation the KKT conditions can be rewritten as the equations (2.5).

**Remark 2.6.** We include an $0 < \epsilon \ll 1$ term to ensure all three regions for the integer program have a non-zero width. The resulting regions are $u_t \leq g_t - \epsilon$, $g_t - \epsilon \leq u_t \leq g_t + \epsilon$, and $u_t \geq g_t + \epsilon$, and note that the binary variables $x_1^t, x_t^2, x_t^3$ indicate if $u_t$ respectively belongs to one of these three regions.

**Remark 2.7.** If $g_t$ is not fixed, as would be the case in an optimization problem for personalizing physical activity goals, then the constraints (2.5) can be further reformulated as MILP constraints using the approach discussed in Section 2.4.5.

### 2.4.4 Exercise Goal Inequalities to Constrain Integer Variables

We define an additional set of inequalities that lead to order of magnitude faster computation times when computing the MLE. Social cognitive theory suggests that if an exercise goal $g_t$ is not achieved at a particular time point $t$ (i.e., $u_t < g_t$), then it will not be achieved at time $t+1$ unless the goal decreases $g_{t+1} < g_t$ or an office visit occurs $d_{t+1} = 1$. This insight leads to additional inequalities on the integer variables.

**Proposition 2.5.** For fixed $g_t$, the logical constraint $(u_t < g_t, g_{t+1} \geq g_t, d_{t+1} = 0) \Rightarrow (u_{t+1} < g_{t+1})$ can be formulated as linear inequalities:

$$\begin{aligned} x^1_{t+1} &\geq x^1_t - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0) \\ x^2_{t+1} &\leq x^2_t + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \\ x^3_{t+1} &\leq x^3_t + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0). \end{aligned} \tag{2.6}$$

The complete proof of this proposition can be found in Appendix A.1, but here we will present some intuition for the proof. The first equation indicates that if $u_t \leq g_t$, then this must also be true at time $t + 1$ unless an office visit is scheduled or the goal has been reduced. The remaining equations indicate that for $u_{t+1} \geq g_{t+1}$ to hold either this condition must have been met at time $t$, an office visit has been scheduled, or the goal has been reduced.

**Remark 2.8.** When $g_t$ is not fixed, the above constraints (2.6) can be further reformulated as MILP constraints using big-M formulations (Vielma, 2015).

These inequalities further constrain the estimates beyond the equations of the utility-maximizing framework. Restated, depending upon the parameters the utility-maximizing framework could potentially predict that goals are not attained at $t$ but then attained at $t + 1$ because of an increase in weight $w_{t+1} > w_t$. We constrain the parameters using the inequalities (2.6) so as to prevent such behavior in the utility-maximizing framework.

### 2.4.5 Addressing Bilinear Terms

Because we are jointly estimating noisy/missing data and parameters, our optimization model contains nonconvex quadratic terms. For instance, the dynamics on $p_t$ (2.3) have the nonconvex quadratic term $\mu \cdot \mathbb{1}(u_t \geq g_t)$. It is difficult to directly solve nonconvex mixed-integer quadratically constrained quadratic programs (MIQCQP) problems, and so we discuss reformulations that allow us to solve the resulting problem more efficiently. We begin by reformulating (2.3).

**Proposition 2.6.** The dynamics on $p_t$ (2.3) can be represented by the linear constraints:

$$\begin{aligned} p_{t+1} &\geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} \\ p_{t+1} &\leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + M \cdot (1 - x^1_t) \\ p_{t+1} &\geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu - M x^1_t \\ p_{t+1} &\leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu. \end{aligned} \tag{2.7}$$

A complete proof of this proposition can be found in Appendix A.1, but here we will present some intuition for the proof. First note that the bilinear term $\mu \cdot \mathbb{1}(u_t \geq g_t)$ in equation (2.3) can be reformulated using the variables $x_t^1, x_t^2, x_t^3$ from the integer reformulation of the KKT conditions (2.5) as $\mu(1 - x_t^1)$. Then, since the resulting term is a product of a continuous and binary variable it can be linearized using standard techniques resulting in the constraint set (2.7).

Finally, to eliminate bilinear terms in our MLE formulation note that the exact-linearization dynamics of $p_t$ (2.7) have the term $\gamma \cdot p_t$, the dynamics of $s_t$ (2.2) have the term $\gamma \cdot (s_t - s_0)$, and the integer-reformulated KKT conditions for decision-making with goals (2.5) have the terms $2r_u u_t, 2r_f f_t$. When we fix the value of $\gamma, r_f, r_u$, the resulting MLE formulation will be a MILP. We use an enumeration approach, as described in the next subsection, to address these final bilinear terms.

### 2.4.6 MILP Formulation of MLE

We reformulate the initial MLE problem $\mathbf{P_{mle}}$ as optimization problem described below, $\mathbf{P_{mle-milp}}$. This is a MILP for fixed values of $\gamma, r_f, r_u$ and after rewriting the absolute values using linear constraints (as in Section 6.1.1 of (Boyd and Vandenberghe, 2004), for example), because $a, b, r_f, r_u, \gamma$ are constants when solving $\mathbf{P_{mle-milp}}$. The full MILP formulation for MLE can be found in the A.3.

$$\begin{array}{ll} \min & \sigma_1^{-1/2} \sum_{i=1}^{n_w} |\tilde{w}_{t_i} - w_{t_i}| + \sigma_2^{-1/2} \sum_{i=1}^{n_u} |\tilde{u}_{\tau_i} - u_{\tau_i}| + \sigma_3^{-1/2} \sum_{t=1}^{n} |z_t| \\ \text{s.t.} & (2.1), (2.4), \text{ for } t = 1, \dots, m-1 \\ & (2.1), (2.2), (2.5), (2.6), (2.7), \text{ for } t = m, \dots, n. \end{array} \qquad \mathbf{P_{mle-milp}}$$

Recall that (2.1) are weight dynamics, (2.2) are dynamics on the $s_t$ parameter, (2.4) are KKT conditions for the decision-making model without goals, (2.5) are integer-reformulated KKT conditions for the decision-making model with goals, (2.6) are the exercise goal inequalities that constrain the integer variables, and (2.7) are exact-linearization dynamics of $p_t$.

If $\gamma, r_f, r_u$ are not fixed, then $\mathbf{P_{mle-milp}}$ is a nonconvex MIQCQP. To solve $\mathbf{P_{mle-milp}}$, observe that we can enumerate over $\gamma, r_f, r_u$ and solve a series of MILP's. This is computationally viable because we only need to enumerate over three variables. We can gain an additional computational speedup by using a simple and accurate approximation that allows us to compute the MLE by solving a single MILP. The approximation is due to an observation we made while using enumeration to compute the MLE. We noticed that the MLE was insensitive to the values of $\gamma, r_f, r_u$: There was less than a 5% difference in the objective value over a large range of values for $\gamma \in [0.8, 1]$ and $r_f, r_u \in [1 \times 10^{-7}, 1 \times 10^{-5}]$, and the estimates of the type parameters were relatively constant over this range as well. As a result, we approximate this problem by fixing $\gamma = 0.85$, $r_f = 8.1633 \times 10^{-6}$, and $r_u = 1 \times 10^{-6}$ for all individuals: This allows us to compute the MLE by solving $\mathbf{P_{mle-milp}}$ for a single value of $\gamma, r$, which is a single MILP. This approximation also reduces the number of parameters we are trying to estimate.

## 2.5 Bayesian Predictions of Individual Trajectories

Problem $\mathbf{P_{mle-milp}}$ provides joint estimation of noisy/missing data and model parameters. However, this is not by itself useful for predicting the future weight loss trajectory of an individual given a short period of initial data. We would ideally like a framework to provide such predictions under different intervention scenarios, since this would support the adaptive design of personalized interventions. Moreover, we would like the predictions to be able to leverage past/historical data in order to improve the accuracy of predictions. Given this last constraint, a natural choice for predictions is to use this past data for a prior distribution in a Bayesian framework.

In particular, suppose we have past/historical data from many individuals that have completed the entire weight loss intervention. We can perform MLE to estimate the parameters for the utility-maximizing framework for each of these individuals. Then we can form our priors by computing histograms of these estimates. Let $t_f$ be the total length of an intervention, and define $\Theta = (\theta_1, \dots, \theta_{t_f})$. We use the pdf notation $\hat{\psi}(\Theta)$ to collectively refer to a set of histograms for the parameters $\Theta$, because these histograms are be assumed to be normalized such that they are a pdf.

Now suppose we have an additional individual that has completed only $T$ days of the intervention and has a remaining $t_f - T$ days left in the intervention, where $t_f$ is the total days in the intervention. The data available for this new individual is $(t_i, \tilde{w}_{t_i})$, for $i = 1, \dots, n_w$, and $(\tau_i, \tilde{u}_{\tau_i})$, for $i = 1, \dots, n_u$, where $n_w$ are the number of weight measurements, $n_u$ are the number of step measurements, and the noise model is as before. We would like to construct an optimization model whose solution provides a prediction of the distribution of the individual's weight at the end of the intervention at time $t_f$ using the histograms of the past individuals and the first $T$ days of data for this new individual. In this section, we demonstrate that we can incorporate data-driven histograms as priors in Bayesian estimation using integer programming.

### 2.5.1 Initial Formulation for Bayesian Estimation

Our goal is to compute $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U})$, which is the *posterior distribution* of weight at the end of the intervention $w_{t_f}$ conditioned (i) on the intervention parameters $C = (d_1, g_1, \dots, d_{t_f}, g_{t_f})$, and (ii) on the data available for the new individual $\tilde{W} =$

$((t_i, \tilde{w}_{t_i}),\ \text{for } i = 1, \ldots, n_w)$ and $\tilde{U} = ((\tau_i, \tilde{w}_{\tau_i}),\ \text{for } i = 1, \ldots, n_u)$. To accomplish this, we apply Bayes's theorem and then eliminate nuisance parameters by averaging over them.

First we calculate $\psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U})$, which is the joint posterior distribution of weight $W = (w_1, \ldots, w_{t_f})$, steps $U = (u_1, \ldots, u_{t_f})$, caloric intake $F = (f_1, \ldots, f_{t_f})$, and type $\Theta$. This requires specifying prior distributions for $W, U, F, \Theta$. The typical approach is to choose priors that admit efficient computation or are uninformative/non-constraining (Gelman et al., 2013). Because we have data from past individuals, we can use the histogram $\hat{\psi}(\Theta)$ as a prior distribution for $\Theta$. We choose a uniform prior distribution for $W, U, F$ because this is relatively uninformative/non-constraining (Gelman et al., 2013). Consequently, applying Bayes's theorem yields $\psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U}) = \frac{1}{Z} \cdot \psi(\tilde{W}, \tilde{U} \mid W, U, F, \Theta, C) \cdot \hat{\psi}(\Theta)$, where $\psi(\tilde{W}, \tilde{U} \mid W, U, F, \Theta, C)$ is the likelihood of the observations conditioned on the parameters of the utility-maximizing framework, $Z$ is a normalization constant that ensures the integral of the posterior is one, and we have used the fact that $\psi(W) = \psi(U) = \psi(F) = 1$ over their supports since they are uniform. Recall that the log-likelihood (i.e., $\log \psi(\tilde{W}, \tilde{U} \mid W, U, F, \Theta, C)$) is given by the objective and constraints of $\mathbf{P_{mle-milp}}$.

The next step is to eliminate nuisance parameters, which can be accomplished in principle by averaging (Gelman et al., 2013). Averaging gives $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U}) = \int \psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U}) \cdot dW_{-t_f} \cdot dU \cdot dF \cdot d\Theta$, where $W_{-t_f} = (w_1, \ldots, w_{t_f - 1})$. However, this integral is difficult to compute both symbolically (because of integer constraints in the formulation of the model) and computationally (the posterior $\psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U})$ can be sharply peaked and so Monte Carlo-based approaches converge slowly). (In fact, our initial approach was to use a Monte Carlo algorithm to compute the posterior distribution, but we found through empirical testing that the resulting posterior was simply a uniform distribution with a very broad support, which indicates convergence to the actual posterior was too slow for making accurate predictions with the posterior; such slow convergence is not surprising given the high-dimensionality of the nuisance parameters.) Our approach is to use the profile likelihood (Severini, 1999, Murphy and Vaart, 2000) as an approximation: The profile likelihood is computed by an optimization problem $\mathbf{P_{pl}}$ that is given by $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U}) \approx \max_{W_{-t_f}, U, F, \Theta} \psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U})$, and our approximation can be justified by arguments relating the asymptotic consistency of Bayesian and MLE estimation under general conditions (Severini and Wong, 1992, Severini, 1999, Gelman et al., 2013). The key computational question is how to solve $\mathbf{P_{pl}}$. The normalizing factor $Z$ can be computed by numerically integrating a one-dimensional function.

### 2.5.2 Histogram Construction

Before constructing the histograms defining $\hat{\psi}(\Theta)$, we need to specify which parameters are statistically independent. Assuming every parameter is correlated will not be successful because it would require high-dimensional histograms, which will be a statistically poor estimate of the true parameter distribution. Hence, specifying that some parameters are independent will enable expressing $\hat{\psi}(\Theta)$ in terms of low-dimensional histograms. Therefore, we assume that $\mu, q, s_0, \beta_0, \delta_0$ are jointly independent. Furthermore, we assume that $\beta_{k+1}$ conditioned on $\beta_k$ is jointly independent with the other parameters. Similarly, $\delta_{k+1}$ conditioned on $\delta_k$ is assumed to be jointly independent with the other parameters. Lastly, we assume that the conditional relationships between $\beta_{k+1}, \beta_k$ and between $\delta_{k+1}, \delta_k$ are not a function of $k$.

**Remark 2.9.** Under our assumptions, we can factor the histogram as $\hat{\psi}(\Theta) = \hat{\psi}(\mu) \cdot \hat{\psi}(q) \cdot \hat{\psi}(s_0) \cdot \hat{\psi}(\beta_0) \cdot \prod_{k=0}^{n_d} \hat{\psi}(\beta_{k+1} \mid \beta_k) \cdot \hat{\psi}(\delta_0) \cdot \prod_{k=0}^{n_d} \hat{\psi}(\delta_{k+1} \mid \delta_k)$, where $n_d$ be the number of office visits.

It will be the case that the objective function we use will involve the logarithm of $\hat{\psi}(\Theta)$, and so the above remark implies that we can construct a MILP formulation of the resulting optimization problem as long as we are able to define MILP representations of $\hat{\psi}(X), \hat{\psi}(X_{k+1} \mid X_k)$, where $X$ is a random variable. Observe, that these constituent histograms are piecewise constant:

**Remark 2.10.** We can represent the one-dimensional histogram for parameter $X$ (where $X$ could be any of $\mu, q, s_0, \beta_0, \delta$) as $\hat{\psi}(X) = \sum_{i=1}^{m_x} \pi_i^x \cdot \mathbb{1}(h_i^x \leq X \leq h_{i+1}^x)$, where $m_x$ is the number of bins, $h_i^x$ are the edges of these bins, and $\pi_i^x$ is the value of the histogram in the $i$-th bin.

**Remark 2.11.** We can represent the histograms for parameter $X_{k+1}$ conditioned on $X_k$ (where $X$ could be any of $\beta_k, \delta_k$) as $\hat{\psi}(X_{k+1} \mid X_k) = \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \pi_{i,j}^x \cdot \mathbb{1}(h_i^x \leq X_{k+1} \leq h_{i+1}^x) \cdot \mathbb{1}(\phi_j^x \leq X_k \leq \phi_{j+1}^x)$, where $m_x$ is the number of bin divisions in the $X_{k+1}$ dimension, $\eta_x$ is the number of bin divisions in the $X_k$ dimension, $h_i^x$ are the edges of the bins in the $X_{k+1}$ dimension, $\phi_i^x$ are the edges of the bins in the $X_k$ dimension, and $\pi_{i,j}^x$ is the value of the histogram in the $(i, j)$-th bin. Note that the histogram values $\pi_{i,j}^x$ should be normalized such that the above representation is a conditional distribution – an incorrect normalization would cause the above representation to be a joint distribution instead.

### 2.5.3 MILP Formulation for Computing Posterior Distribution of Final Weight

One of our goals is to show that data-driven prior distributions can be used to perform Bayesian estimation by formulating the problem as a MILP. Here, we focus on approximating the posterior $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U})$ by solving $\mathbf{P_{pl}}$. It is worth noting that an almost identical formulation can be used to perform Bayesian *maximum a posteriori* (MAP) estimation with data-driven priors by solving problem $\mathbf{P_{map}}$, which is given by $\max_{W,U,F,\Theta} \psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U})$; compare this problem to $\mathbf{P_{pl}}$. Observe that because a histogram is a piecewise constant function, it can be represented using inequality constraints with integers (Vielma, 2015). This requires some minor reformulations, which we describe below, in order to ensure linearity of the optimization model.

**Proposition 2.7.** The objective of $\mathbf{P_{pl}}$ (after computing its negative logarithm) is

$$\begin{aligned}&\sigma_1^{-1/2} \sum_{i=1}^{n_w} |\tilde{w}_{t_i} - w_{t_i}| + \sigma_2^{-1/2} \sum_{i=1}^{n_u} |\tilde{u}_{\tau_i} - u_{\tau_i}| + \sigma_3^{-1/2} \sum_{t=1}^{n} |z_t| + \\ &2^{-1/2} \sum_{X \in \{\mu, q, s_0, \beta_0, \delta_0\}} \sum_{i=1}^{m_x} \log \pi_i^x \cdot y_i^x + 2^{-1/2} \sum_{X \in \{\beta, \delta\}} \sum_{k=0}^{n_d - 1} \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \log \pi_{i,j}^x \cdot y_{i,j}^{x,k},\end{aligned} \tag{2.8}$$

subject to constraints for one-dimensional histograms

$$\sum_{i=1}^{m_x} h_i^x \cdot y_i^x \le X \le \sum_{i=1}^{m_x} h_{i+1}^x \cdot y_i^x; \quad \textstyle\sum_{i=1}^{m_x} y_i^x = 1; \quad y_i^x \in \{0, 1\}, \ \forall i = 1, \dots, m_x, \tag{2.9}$$

for all $X \in \{\mu, q, s_0, \beta_0, \delta_0\}$, and constraints for conditional histograms

$$\begin{aligned}&\textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} h_{i,j}^x \cdot y_{i,j}^{x,k} \le X_{k+1} \le \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} h_{i+1}^{x,k} \cdot y_{i,j}^x \\ &\textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \phi_{i,j}^x \cdot y_{i,j}^{x,k} \le X_k \le \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \phi_{i+1}^x \cdot y_{i,j}^{x,k} \\ &y_{i,j}^{x,k} \in \{0, 1\}, \ \forall i = 1, \dots, m_x, \ j = 1, \dots, \eta_x; \quad \textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} y_{i,j}^{x,k} = 1,\end{aligned} \tag{2.10}$$

for all $X \in \{\beta, \delta\}$ and $k = 0, \dots, n_d - 1$.

A complete proof of this result can be found in Appendix A.1 but here we will present some intuition for the proof. Taking the log of $\hat{\psi}(\Theta)$ we obtain an objective function which comprises of both terms $\log \psi(\tilde{W}, \tilde{U} \mid W, U, F, \Theta, C)$ as well as the equations for the one and two dimensional histograms. Then by defining indicator variables $y_i^x \in \{0, 1\}$ for $x \in \{\mu, q, s_0, \beta_0, \delta_0\}$, where $y_i^x$ is equal to 1 if parameter $x$ is in interval $i$ of the histogram, and similarly indicator variables $y_{i,j}^{x,k} \in \{0, 1\}$ for parameters $x \in \{\beta, \delta\}$, we can rewrite the objective function as the form above.

**Remark 2.12.** An important benefit of the rewritten objective (2.8) and subsequent constraints (2.9), (2.10) is that they are linear in the decision variables. Note that the $y$ decision variables in these equations are binary variables and indicate which bin of the histogram the corresponding variable belongs to.

The posterior is computed by solving a series of MILPs and then using numerical integration to compute the normalization constant $Z$. In particular, define the following parametric (in $\omega$) MILP:

$$\begin{aligned}
\ell(w_{t_f} = \omega) = \min\ & (2.8) \\
\text{s.t. } & (2.1), (2.4),\ \text{for } t = 1, \dots, m-1 \\
& (2.1), (2.2), (2.5), (2.6), (2.7),\ \text{for } t = m, \dots, n \\
& (2.9),\ \text{for } X \in \{\mu, q, s_0, \beta_0, \delta_0\} \\
& (2.10),\ \text{for } X \in \{\beta, \delta\},\ k = 0, \dots, n_d - 1; \quad w_{t_f} = \omega.
\end{aligned}$$

The complete formulation can be found in the A.4.

Let $\kappa_2 = \min_i \ell(w_{t_f} = \omega_i)$. If we solve $\mathbf{P_{pl-milp}}$ over a grid of values $\omega_1, \dots, \omega_{n_g}$, then we can compute the normalization $Z$ by numerically integrating the set of points $(\omega_i,\ \exp(-\ell(w_{t_f} = \omega_i) + \kappa_2))$, for $i = 1, \dots, n_g$, where (i) we take the exponent of the negative of $\ell(\cdot)$ because we reformulated the objective for our MILP using a negative logarithm, and (ii) we scale this exponent using $\kappa_2$ because this improves the numerics of the computations. Finally, the posterior at $\omega_i$ is given by $\psi(w_{t_f} = \omega_i \mid C, \tilde{W}, \tilde{U}) = \exp(-\ell(w_{t_f} = \omega_i) + \kappa_2)/Z$. Consequently, we can approximate the posterior distribution of $w_{t_f}$ by solving a series of problem $\mathbf{P_{pl-milp}}$. Observe that in this approximation process, we are in fact approximating the posterior likelihood of the final weight $w_{t_f} = \omega$ at different $\omega$ using different patient behaviors trajectories. Such an approximation approach has been previously proposed and is well-behaved asymptotically as more data is collected (Lindley, 1961, Tierney and Kadane, 1986, Evans and Swartz, 1995). The intuition from Evans and Swartz (1995) for why such an approximation is justified begins with the defining integral $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U}) = \int \psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U}) \cdot dW_{-t_f} \cdot dU \cdot dF \cdot d\Theta$. For a fixed $w_{t_f}$, by the law of large numbers most of the mass of $\psi(W, U, F, \Theta \mid C, \tilde{W}, \tilde{U})$ is concentrated about its maximizer, which corresponds to the minimizer of $\mathbf{P_{pl-milp}}$. Hence we can approximate this integral by considering its behavior at the optimizer. We will further discuss the theoretical properties of this approximation in Chapter 3.

## 2.6 Computational Results and Validation of Predictive Modeling

In this section, we first describe the data source used for the computational results and validation of our predictive model. Next, we provide computational results of solving $\mathbf{P_{mle-milp}}$ to compute MLE and of solving $\mathbf{P_{pl-milp}}$ to compute the Bayesian predictive model. Representative plots are shown in these first two subsections. Cross-validation (Hastie et al., 2009) is used to validate our approach through comparison to a benchmark approach from machine learning, and we specifically consider the prediction of 5% weight loss at 5 months based on the first 30 days of an individual's data. This validation compares all individuals in the data set. We conclude by demonstrating the ability of our approach to make predictions on the weight loss trajectory of an individual as the number of counseling sessions is changed, and we discuss how this can be used for optimization.

### 2.6.1 Data Source of Mobile Phone Delivered Diabetes Prevention Program (mDPP) Trial

We used data from the mDPP trial (Fukuoka et al., 2015), which was a randomized controlled trial (RCT) to evaluate the efficacy of a 5-month mobile phone-based weight loss intervention among overweight English-speaking adults at risk for developing T2DM. The intervention was adapted from the Diabetes Prevention Program (DPP) 2002, 2009, but the frequency of in-person sessions was reduced from 16 to 6 sessions and group exercise sessions were replaced with a home based exercise program to reduce costs. Sixty-one overweight adults were randomized to an active control (accelerometer only) ($n = 31$) group or an mDPP mobile app plus accelerometer intervention ($n = 30$). Demographics are available in (Fukuoka et al., 2015), and changes in primary and secondary outcomes were promising: The intervention group lost an average of $6.2 \pm 5.9$ kg ($-6.8\% \pm 5.7\%$) between baseline and 5-month follow-up compared to the control group's gain of $0.3 \pm 3.0$ kg ($0.3\% \pm 5.7\%$) ($p < 0.001$). The intervention group's steps per day increased by $2551 \pm 4712$ compared to the control's group decrease of $734 \pm 3308$ steps per day ($p < 0.001$).

The data available from this RCT matches that described in Section 2.2. Specifically, we have step data from a digital accelerometer and body weight data recorded at least twice a week every week into the mobile app. We also have access to the age, gender, and height of each individual. After an initial two week period, exercise goals in desired number of steps per day were provided to each individual. The goals increased by 20% each week, starting at 1.2 times the average number of steps during the initial two weeks; the goals increased to a maximum of 12,000 steps a day (about 6 miles of walking). Individuals were also asked to make office visits (at 2, 4, 6, 10, 14, 18, and 20 weeks) during which they received behavioral counseling about their nutritional choices and physical activity.

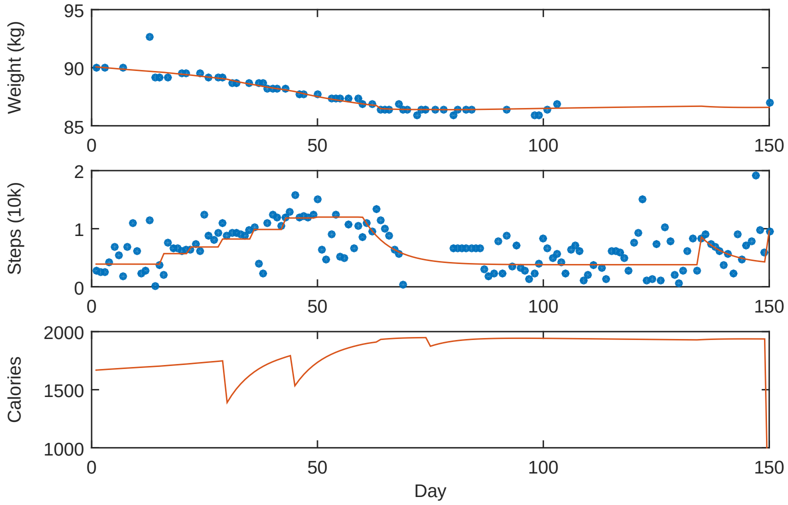


Figure 2.1: Comparison of data (blue dots) with MLE estimates of weight, exercise, and caloric intake (red line).

### 2.6.2 Computational Results

We used the Gurobi solver (Gurobi Optimization, 2015) to solve $\mathbf{P}_{\mathbf{mle-milp}}$ and $\mathbf{P}_{\mathbf{pl-milp}}$. The CVX toolbox (Grant and Boyd, 2014) for MATLAB was used to generate each instance of the MILP. A 2.5GHz laptop computer with 4Gb of RAM was used to generate these results.

#### Results of MLE for Utility-Maximizing Model

The problem $\mathbf{P}_{\mathbf{mle-milp}}$ was solved for each individual in the mDPP. The fastest computation time was 3 sec, the slowest computation time was 550 sec, and the median computation time was 10 sec. The second and third quartiles of computation time were 6 sec and 70 sec, respectively. Overall, the computation was quick and can be easily parallelized because each MILP is solved independently.

Figure 2.1 shows a representative example of the weight, steps, and caloric intake trajectory estimated by solving $\mathbf{P}_{\mathbf{mle-milp}}$. The blue dots are measured data, and the red lines are estimated trajectories. The utility-maximizing framework captures increasing positive impacts from achieving exercise goals, as well as negative impacts from not meeting goals. The MLE reduces noise in measured data and estimates values for time points without data. Observe that the large drops in caloric intake correspond to reductions in the preference of caloric consumption $s_t$ that occurs after an office visit; however, the reductions are not constant for each office visit. This is because the impact of an office visit is characterized by $\beta_t, \delta_t$, which are random variables. Moreover, when we computed the conditional histograms for $\beta_{t+1}$ given $\beta_t$ and for $\delta_{t+1}$ given $\delta_t$, we empirically found that these histograms were such that they indicated subsequent office visits are generally less effective in encouraging increases in physical activity and reductions in caloric intake.

From a clinical standpoint, an additional benefit of our utility-maximizing framework is its ability to estimate caloric intake. Effective mobile technologies for directly measuring caloric intake are not commercially available, and self-reported caloric intake diaries are known to be highly inaccurate (Schoeller et al., 1990). Our approach indirectly estimates this by integrating physiology into the framework. This can be used to improve self-monitoring of an individual's food consumption.

### Results of Bayesian Trajectory Prediction using MILP Formulation

Problem $\mathbf{P_{pl-milp}}$ was solved using the first month of data for each individual in the intervention group of the mDPP in order to compute a posterior distribution of $w_{t_f}$. To generate the histograms for $\mathbf{P_{pl-milp}}$, we used the MLE parameters for the remaining individuals computed using the entire data set for these individuals. We did *not* use an individual's data when computing the histogram used to make predictions for that particular individual; we constructed a different histogram for each individual by using the data excluding that individual.

For our computations, we chose $n_g = 100$ grid points at which we computed the posterior. The fastest, slowest, and median computation times were 190 sec, 1000 sec, and 360 sec, respectively. The second and third quartiles of computation time were 230 sec and 470 sec, respectively. Overall, the computation was relatively quick and can be easily parallelized because each MILP is solved independently.

A representative example of the posterior likelihood $\psi(w_{t_f} \mid C, \tilde{W}, \tilde{U})$ for the final weight of an individual (at 5 months) conditioned on 1 month of weight and step data is shown in Figure 2.2. The dashed line denotes the initial weight of the individual before starting the weight loss intervention, and the dotted line represents a final weight corresponding to 5% weight loss. We can also plot the entire weight, exercise, and caloric intake trajectories corresponding to the MAP estimate: This is shown in Figure 2.3. Data from the first month (dark blue and left of the dotted line) was used to compute the posterior and the MAP estimate of the past and future trajectories. The MAP prediction of the future trajectories is compared to the actual measurements (light blue and right of the dotted line); there is good agreement between the predicted and actual weight trajectories. An additional benefit of this approach is its ability to estimate past caloric intake.

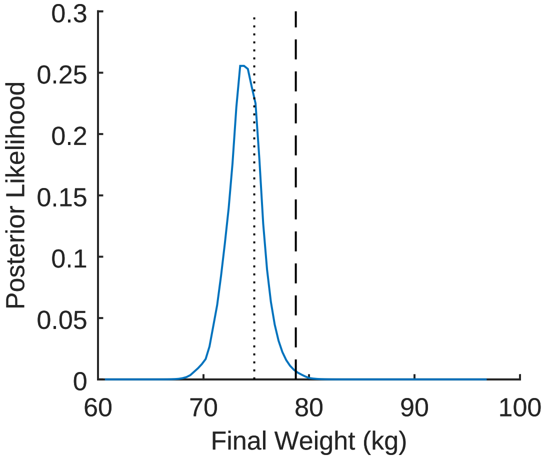


Figure 2.2: Posterior likelihood of final weight conditioned on 30 days of data (solid) compared to initial weight (dashed) and final weight corresponding to a 5% weight loss (dotted).

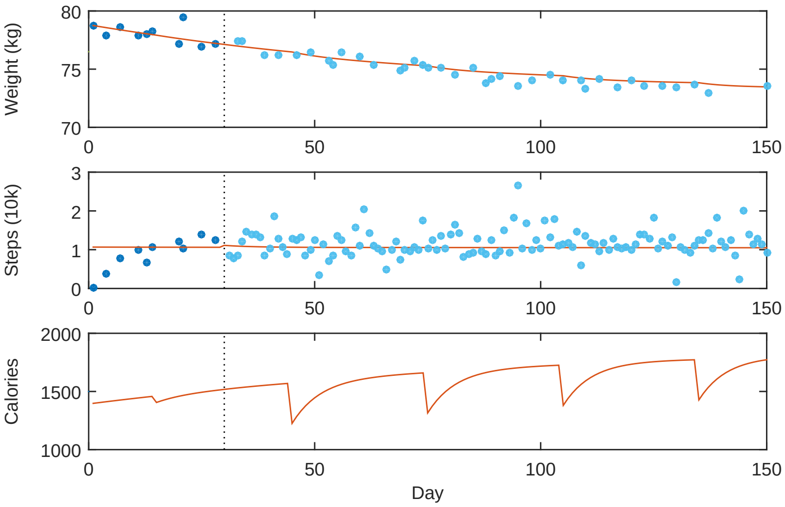


Figure 2.3: Comparison of MAP estimates of weight, exercise, and caloric intake trajectories (trained using data marked with dark blue dots) with future data not used to computed estimates (light blue dots).

### 2.6.3 Predicting Clinically Significant Weight Loss

This subsection evaluates the ability of our predictive model from Section 2.5 to predict whether an individual will achieve clinically significant weight loss at the end of the intervention. We refer to a situation where an individual achieves 5% weight loss as a *positive*, and similarly if an individual does not achieve 5% weight loss then this is be a *negative*. We validate the predictive capabilities of our model by comparing it to three standard methods from machine learning. Specifically we consider a linear support vector machine (SVM) model, a decision tree model, and a logistic regression model for classification (Hastie et al., 2009). We additionally consider a version of our predictive model that does not incorporate a Bayesian prior in order to validate that our Bayesian approach improves prediction accuracy. For the purpose of comparison, we specifically consider a scenario in which the first 30 days of mobile phone data are used to predict whether an individual will achieve 5% weight loss after 5 months of participating in the weight loss intervention. Cross-validation (Hastie et al., 2009) is used to separate the data into a training set that is used to estimate the models and a hold-out set that is used to quantitatively validate the model.

#### Machine Learning Models

Let $x \in \mathbb{R}^2$ be a vector of percent weight loss to date and percent of step goals met, and let $y$ be such that if $y = 1$ then an individual has achieved at least 5% weight loss and $y = -1$ otherwise. Machine learning methods use data in this form to fit functions $f : \mathbb{R}^2 \to \{-1, 1\}$ to best capture the relationship between $x$ and $y$. We refer to the output of this function as $\hat{y}(x) = f(x)$, to signify that we are generating an estimate of the $y$ values. The value $\hat{y}(x) = -1$ is a prediction that the individual *will not* achieve 5% weight loss after 5 months, and $\hat{y}(x) = +1$ is a prediction that the individual *will* achieve 5% weight loss after 5 months.

A *linear SVM* is the predictive model $\hat{y}(x) = \text{sign}(\beta_0 + x'\beta)$. The hyperplane $\beta_0 + x'\beta$ cuts the space $\mathbb{R}^p$ into two regions, and the two sides of the hyperplane are predicted to be positive or negative, respectively. The parameters $\beta_0, \beta$ are computed by a quadratic program (Hastie et al., 2009), and we used the MATLAB Statistics and Machine Learning Toolbox to identify the SVM parameters using data from the mDPP trial. The identified parameters are 64 for percent weight loss to date and 1.715 for percent of step goals met; the parameter values normalized by sample standard deviation were similar. These magnitudes indicate that for predicting 5% weight loss: percent weight loss to date is the most important feature and percent of step goals met is the second most important. Because all parameters are positive, this means increased weight loss to date and percent of step goals met both lead to increased likelihood of achieving 5% weight loss.

A *decision tree* model (i.e., classification and regression trees or CART) is a sequential classifier. Each node of the tree partitions a different column of the data to ensure maximum separation between the two classes, and each leaf of the tree is assigned a label of 1 or $-1$. For prediction, data is compared along the nodes of the tree and then assigned a value that corresponds to the leaf of the final comparison. Computing the optimal decision tree model is NP-hard, and heuristics are used to construct these models (Hastie et al., 2009). For our implementation, we used the MATLAB Statistics and Machine Learning Toolbox to train the decision tree model from the mDPP data. Our trained decision tree model first branches on the percent of weight lost to date, with those who lost at least 2.6% being classified to the class which will achieve 5% weight loss. Next, the model branches on the average amount of exercise goals met, with those who met at least 84% of their goals being classified as successful and the remainder as unsuccessful.

A *Logistic Regression* model specifies a classifier of the form $\hat{y}(x) = 2\cdot\mathbf{1}\{\frac{1}{1+\exp(-\beta_0 - x'\beta)} \geq \frac{1}{2}\} - 1$. This probabilistic interpretation of this classifier is that the labels transformed to $\{0, 1\}$ follow a Bernoulli distribution with parameter $p$ where $\log(\frac{p}{1-p}) = \beta_0 + x'\beta$. Hence if the probability that $y = 1$ is greater than 0.5 we predict $\hat{y} = 1$, and otherwise we predict $\hat{y} = -1$. The problem of training a logistic regression model can be posed as a convex optimization problem (Hastie et al., 2009) that can be solved by stochastic gradient descent. For our analysis, we used the MATLAB Statistics and Machine Learning Toolbox to train the logistic regression model. The coefficient for weight lost to date was 73 and the coefficient for percent of goals met was 0.889. These coefficients are similar to the SVM coefficients, which is unsurprising since logistic regression can be interpreted as a continuous relaxation of linear SVM (Hastie et al., 2009).

### Adjusting True and False Positive Rate of Predictions

The quality of our models can be evaluated by estimating and comparing the true and false positive rates of different models. The *true positive rate* (TPR) specifies the probability of a model correctly predicting a positive, and the *false positive rate* (FPR) quantifies the probability of a model incorrectly predicting a positive. In making predictions, there is tradeoff between the TPR and FPR. It is customary for practitioners to choose the FPR, and this choice fixes the TPR (Bickel and Doksum, 2006, Lehmann and Romano, 2006). Choosing the FPR requires an understanding of how the model is used to make predictions and how parameters in the model impact the FPR. For instance, we can adjust the FPR of a linear SVM model by choosing the value of $\beta_0$. For example, if $\beta_0 = -\infty$, then the prediction will always be $-1$; similarly, if $\beta_0 = +\infty$, then the prediction will always be $+1$. By choosing intermediate values for $\beta_0$, we can adjust the FPR of the model. To specify the FPR of the Bayesian predictive model, we compute the posterior probability of 5% weight loss

$\mathbb{P}(w_{t_f} \leq 0.95w_0 \mid C, \tilde{W}, \tilde{U}) = \int_{-\infty}^{0.95w_0} \psi(w_{t_f} \mid C, \tilde{W}, \tilde{U}) \cdot dw_{t_f}$ and then threshold this at successively lower levels. This is similar to the standard approach used to choose the FPR for logistic regression.

### Estimating an ROC Curve

It is common to choose the FPR using a receiver operating characteristic (ROC) curve. An ROC curve explicitly displays the tradeoff between the TPR and FPR. We can estimate such a curve for the various machine learning models and our Bayesian predictive model both with and without the empirical prior distribution. In particular, we use leave-one-out cross-validation (Hastie et al., 2009) to estimate each ROC curve. The idea of this standard approach is that when making the prediction for each individual, we use a model that was computed using data from everyone excluding the present individual. The final result is a summation over the predictions for each individual. The benefit of this approach is we do not use data from a specific individual when making the prediction for that specific individual.

We estimated an ROC curve for each of the models using leave-one-out cross-validation, and these ROC curves are shown in Figure 2.4. These ROC curves compare the prediction accuracy for all individuals. The ROC curves have been smoothed using a binormal model (Metz et al., 1998), and the unsmoothed version of the ROC curves can be found in Appendix A.5. The results show that our predictive modeling framework is competitive in terms of prediction accuracy with the linear SVM, logistic regression, and decision tree models, which further justifies our choice of the utility-maximizing framework and its ability to capture "irrational" discounting in the decision-making of individuals participating in the intervention. Furthermore, our predictive model with the Bayesian empirical prior makes slightly better predictions than our predictive model without a Bayesian prior, though the difference in their ROC curves is not statistically significant ($P = 0.16$) when compared using a standard hypothesis testing approach developed by Hanley and McNeil (1983). In contrast, the difference in the ROC curves of our predictive model (with and without the prior) and the benchmark approaches of linear SVM, logistic regression, and decision tree models is statistically significant ($P = 0.001$). Our empirical results suggest that the Bayesian prior gives a slight improvement for this data set, but this is not expected to generally hold. Essentially, we expect that using a prior will give improvements in prediction accuracy when an individual is similar to those individuals used to construct the prior. On the other had, if an individual is very different from those used to construct the prior, then we expect the prior to make predictions worse. However, the situation may be improved with a demographics-dependent prior: We could imagine constructing different priors for individuals with different demographics. Then when making predictions for an individual, we could either use a prior constructed by the data of those

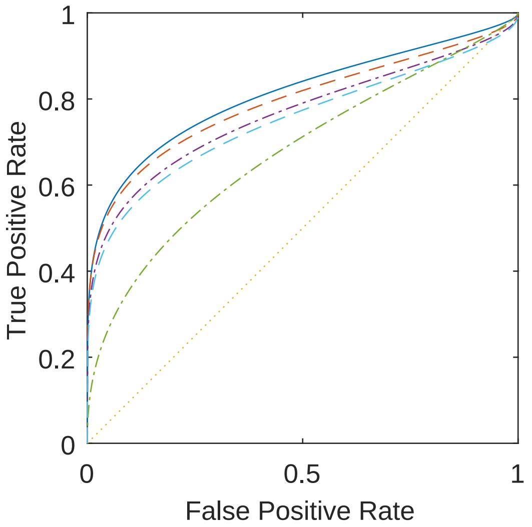


Figure 2.4: ROC curves computed using leave-one-out cross-validation for our predictive model with an empirical Bayesian prior (blue solid), our predictive model without a Bayesian prior (red dashed), linear SVM model (purple dash dot), decision tree model (green dashed dot), and logistic regression (cyan dashed) are compared.

with matching demographics, or not use a prior if the individual has very different demographics than was used to construct any of the priors.

### 2.6.4 Personalizing Goal Setting Using the Predictive Model

One of our reasons for developing a predictive model is to enable the design of approaches for optimizing elements of large weight loss programs. In contrast to other predictive models, our behavioral framework can be used to formulate an optimization problem to determine the number of visits, timing of visits, and the physical activity goals for each individual in order to maximize the expected number of individuals that achieve clinically significant weight loss at the end of the program. It is in this way that our predictive model has the potential to be used to personalize the weight loss program for each individual.

Here, we present an example that demonstrates the ability of our model to make predictions about how future weight loss changes as the step goals for an individual are changed. Figure 2.5 shows the posterior likelihood of final weight of an individual conditioned on 50 days of data and on either having 12,000 steps/day goals after 50 days (dash dotted) or having 8,000 steps/day goals after 50 days (solid). When the goals are 8,000 steps/day, our model predicts a 51% chance of achieving 5% weight loss

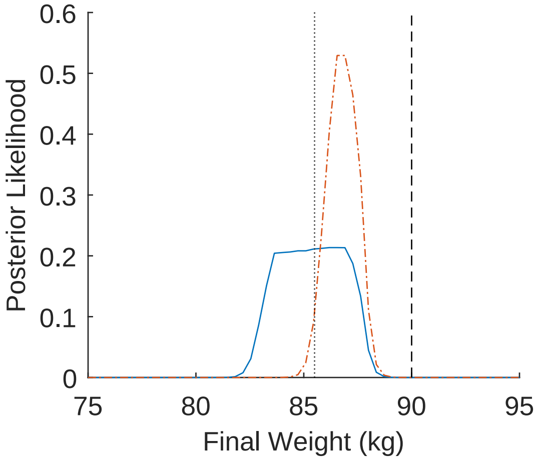


Figure 2.5: Posterior likelihood of final weight of an individual conditioned on 50 days of data and conditioned on either having 12,000 steps/day goals after 50 days (dash dotted) or 8,000 steps/day goals after 50 days (solid), and compared to initial weight (dashed) and final weight corresponding to a 5% weight loss (dotted).

and that the expected final weight conditioned on not achieving 5% weight loss is 86.6 kg. When goals are 12,000 steps/day, our model predicts a 3% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 86.8 kg. Our model predicts 8,000 steps/day goals are superior to 12,000 steps/day goals for motivating this individual to increase their physical activity and consequently lose weight.

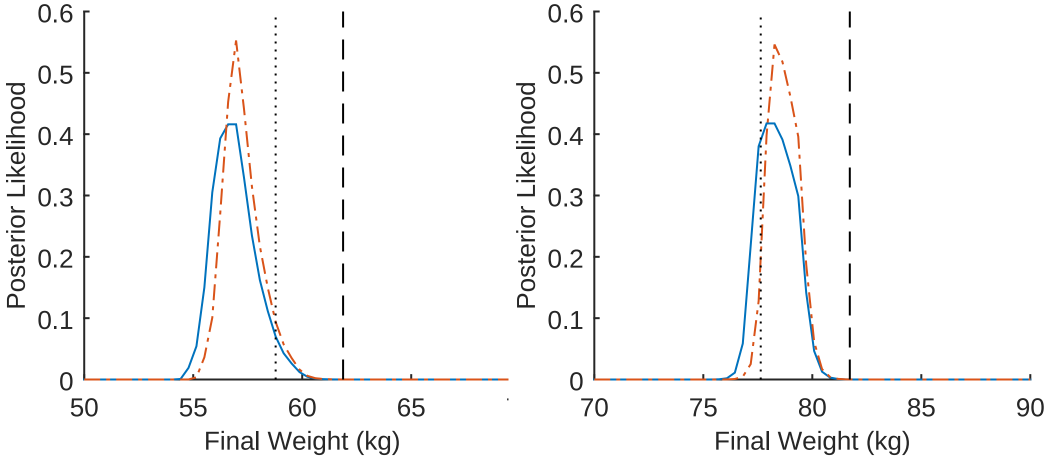
The above example shows the possibility of improving weight loss outcomes, and next we briefly describe how an optimization model can be constructed (based on our predictive model) to personalize the weight loss program. We will discuss the full details of these optimization models in Chapter 3, and we have completed a rigorous clinical trial to experimentally validate the efficacy of our predictive modeling and optimization framework in a setting where the aim was to only increase the physical activity of individuals through a mobile phone app (Zhou et al., 2018). Specifically, the weight loss program can be personalized by first solving our formulation for computing the MAP estimate of an individual's type. Next, we solve the problem $\min w_{t_f}$ subject to constraints defined by our predictive model and the MAP estimate of the individual's type; this problem can be written as a MILP using reformulation techniques similar to the ones described in this chapter.

### 2.6.5 Reducing the Number of Office Visits

Most of the costs and person hours spent on administering weight loss programs are associated with conducting office visits. Thus, it is essential to be able to optimize the total number of visits and when they are scheduled. Our model is able to capture differences in predicted weight loss trajectories that occur when changing the number of office visits. For instance, Figure 2.6a shows the posterior likelihood of final weight of an individual conditioned on 50 days of data and on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid). When scheduling 4 office visits, our model predicts a 96% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 59.3 kg. When scheduling 0 office visits, our model predicts a 94% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 59.3 kg. Our model predicts that for this individual the benefit of scheduling additional office visits is minor.

Another example is shown in Figure 2.6b, which displays the posterior likelihood of final weight of another individual conditioned on 50 days of data and on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid). When scheduling 4 office visits, our model predicts an 18% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 78.6 kg. When scheduling 0 office visits, our model predicts a 3% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 78.7 kg. Our model predicts a clinically significant benefit of scheduling additional office visits for this particular individual.

These two examples demonstrate the ability of our predictive model to identify which individuals are responsive to office visits, and thus our predictive model can be combined with an optimization model to reduce the average number of office visits when considering a large number of individuals participating in a weight loss program. We briefly describe how an optimization model can be constructed (based on our predictive model) to reduce the average number of office visits; full details of the corresponding optimization models are out of the scope of this chapter but will be discussed in Chapter 3. Specifically, we can use a decomposition scheme: In the first step of the scheme, we vary the total number of office visits for each individual over a range of values, and we solve the problem $\min w_{t_f}$ subject to constraints defined by our predictive model, the MAP estimate of the individual's type, and the total number of office visits. This problem can be written as a MILP using reformulation techniques similar to the ones described in this chapter. In the second step of the scheme, we solve a knapsack-like problem that allocates the number of office visits to each individual based on the predicted effectiveness of different numbers of office visits.



(a) Individual A

(b) Individual B

Figure 2.6: Posterior likelihood of final weight of an individual conditioned on 50 days of data and conditioned on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid), and compared to initial weight (dashed) and final weight corresponding to a 5% weight loss (dotted).

## 2.7 Conclusion

We constructed a predictive model of individual behavior in a weight loss intervention, employing a utility-maximizing framework based on qualitative concepts from social cognitive theory. MILP formulations were developed to compute (i) parameters of the framework using MLE, and (ii) a Bayesian predictive model using an empirical histogram (constructed using parameters estimated by MLE) as a prior. Model prediction quality was assessed using leave-one-out cross-validation to compute an ROC curve, and the results show that the utility-maximizing framework leads to predictions on par with predictions of a linear SVM model. We concluded by showing how our predictive model is able to capture differences in weight outcomes as the number of office visits is varied, and we briefly discussed how these models may be used in designing algorithms to optimize and personalize office visit schedules and exercise goals.

# Chapter 3

# Interventions with Costly and Infrequent Decisions

## 3.1 Introduction

The increasing availability of data presents an opportunity to transform the design of incentives (i.e., costly inputs that are provided to agents to modify their behavior and decisions) from a single analysis into an adaptive and dynamic process whereby the incentive design is optimized as new data becomes available. Historically, this adaptive setting has been studied under the framework of repeated games (Radner, 1985, Fudenberg et al., 1994, Laffont and Martimort, 2002), where researchers have focused on the analysis and identification of structural properties of effective policies, and on equilibria. In contrast, continuing advances in optimization software and statistical estimation tools, utilized with the vast amount of data now available in many settings, enable a new approach that in many circumstances has the potential to lead to practical tools for designing effective incentives in real-world settings. This approach, which we call *behavioral analytics*, is built around a three step framework: first, we develop a behavioral model that describes the decision-making process of an agent; next, we iterate repeatedly over two steps as new information is collected. In the second step, we use data to estimate behavioral model parameters for each agent and then use these estimates to predict future decisions of each agent; and in the third, we use the estimated behavioral model parameters to optimize a set of costly incentives to provide to each agent. In this chapter, we describe a specific set of tools, models, and approaches that fit into this framework, and that adapt models and incentives as new information is collected while the second and third steps of the framework are repeated.

Specifically, we consider the following discrete-time setting: There is a large pool of agents each with a set of utility function parameters (which we will refer to as the motivational state) and system state at time $t$, and each agent makes a decision at $t$ by maximizing a myopic utility function. A single coordinator makes noisy observations at $t$ of the system states and decisions of each agent, and then assigns behavioral or financial incentives (e.g., bonuses, payments, behavioral goals, counseling sessions) at $t$ to a subset of agents. The incentives change the motivational and system states of assigned agents at time $t+1$, while the motivational and system states of non-assigned agents evolve at $t+1$ according to some dynamics. This process repeats, and time $t$ advances towards infinity in unit increments. Here, the coordinator's problem is to decide what incentives to provide to which agents in order to minimize the coordinator's loss function, a function that depends on the system states and decisions of all agents. This problem is challenging because the motivational states of agents are neither known nor measured by the coordinator, because agents make decisions by maximizing an unknown utility function, because measurements are noisy, and because the coordinator has a fixed budget (over a specified time horizon) from which to allocate incentives.

### 3.1.1 Potential Applications for Behavioral Analytics

The setting described above is found in many domains, including personalized healthcare, demand response programs, and franchise logistics. Below, we elaborate on these potential applications of our framework. The first application is the design of a weight loss program. The coordinator is a clinician and the agents are individuals trying to lose body weight. The next application is the design of a demand response program in which the coordinator is an electric utility company and the agents are homeowners (since they consume electricity). The final application is in decentralized supply chain management where a manufacturer is a coordinator for a large network of independently operated retailers that act as agents.

#### Weight Loss Programs

In a clinically-supervised weight loss program, a clinician provides two types of behavioral incentives to a group of individuals who are trying to lose body weight. The first type of incentive is behavioral goals provided to each individual by the clinician, and it is costless when communication costs are negligible, as is the case with mobile phone-delivered programs (Fukuoka et al., 2015). The second type of incentive is that the clinician can provide a limited amount of counseling to individuals, but this is costly and the clinician must decide how to allocate a limited budget of counseling sessions to the entire pool of individuals. For example, the Diabetes Prevention Program Research Group (2002, 2003, 2009) has shown that such programs lead to a clinically significant loss of 5-7% body weight on average, which can prevent or delay the onset of type 2 diabetes with few side effects. However, these programs are difficult to design because variations in individual motivational states mean there is not just one set of optimal behavioral goals and assignment of counseling sessions, but rather that behavioral goals and the number/timing of counseling needs to be personalized to individuals' motivational states to maximize weight loss.

Personalizing the behavioral incentives for each individual can improve efficacy of weight loss programs and reduce the associated program costs through a reduction in the average amount of counseling for each individual. Mobile phone technology is one promising avenue for implementing such personalization, due to its relatively low cost and pervasiveness among diverse communities (Lopez et al., 2013). Mobile phones allow clinicians to collect real time health data through use of personal logs and devices such as accelerometers, which provides noisy measurements of the health state and decisions of each individual. Randomized controlled trials (RCT's) have found that the use of mobile phones can reduce the cost of implementing weight loss programs with maintaining efficacy (Fukuoka et al., 2015); however, little research to date has explored how to use the data generated by mobile phones and digital accelerometers in order to personalize behavioral incentives (Fukuoka et al., 2011, O'Reilly and Spruijt-Metz, 2013, Azar et al., 2013, Pagoto et al., 2013).

Several adaptive methods have been proposed for designing personalized healthcare treatments, including: multi-armed bandits (Negoescu et al., 2014, Deo et al., 2013, Bastani and Bayati, 2015a), robust optimization (Bertsimas and O'Hair, 2013), and dynamic programming (Engineer et al., 2009). One common approach for optimal treatment design and clinical appointment scheduling has been Markov decision process (MDP) models (Ayer et al., 2015, Mason et al., 2013, Deo et al., 2013, Kucukyazici et al., 2011, Leff et al., 1986, Liu et al., 2010, Wang and Gupta, 2011, Gupta and Wang, 2008, Savelsbergh and Smilowitz, 2016). These methods are designed for situations with infrequent data collection (e.g., only during clinical visits), whereas in weight loss programs the data is collected daily (or more often) using mobile devices. Our work develops an approach that can leverage this increased data availability to better design incentives. Moreover, existing approaches focus either on motivational states characterizing adherence (Mason et al., 2013) or health states describing prognosis (Ayer et al., 2012, Deo et al., 2013, Helm et al., 2015, Wu et al., 2013, Negoescu et al., 2014, Engineer et al., 2009). In contrast, we seek to combine motivational and health states into a single predictive model that is used for personalizing the weight loss program.

### Demand Response (DR)

DR programs are used by electric utilities to alter homeowners' electricity usage to better match electricity generation and reduce peak electricity demand. Utilities incentivize homeowners to shift or reduce electricity consumption using price-based programs (e.g., time-of-day electricity rates). Utilities also incentivize reduced electricity consumption through exchange programs in which a homeowner's inefficient appliances are replaced (for free by the electric utility) with efficient appliances (Palensky and Dietrich, 2011, Deng et al., 2015). Implementations of such DR programs have decreased peak electricity demand by almost 10% and have improved the balance between electricity supply and demand (Lee et al., 2014). However, adverse selection is a major issue in these programs because incentives are often provided to homeowners who already have low electricity consumption or already had plans to replace inefficient appliances.

Better targeting in a DR program may lead to improved efficacy with lower associated costs. For example, electric utilities have the capability to send an auditor to homes to assess what appliance upgrades are needed (PG&E, 2016). Consequently, an electric utility would be interested in finding the most effective way to schedule its auditors and set its rebates and tariffs. Homeowner electricity usage data can be collected by the utility in real time using smart electricity meters, and the adoption rate of these smart electricity meters is increasing in the US (Lee et al., 2014). Moreover, the two way communication capabilities of smart electricity meters and mobile phones can be used to communicate billing and incentive information to homeowners (Darby, 2010), which opens the possibility for better targeting of price-based programs and appliance-replacement programs.

DR programs are often designed using game-theoretic approaches (Saghezchi et al., 2015, Samadi et al., 2010, 2012), multi-armed bandits (Wijaya et al., 2013), convex optimization (Li et al., 2011, Mohsenian-Rad and Leon-Garcia, 2010, Ratliff et al., 2014), dynamic programming (Jiang and Low, 2011, Costa and Kariniotakis, 2007, Molderink et al., 2010), and MDP's (O'Neill et al., 2010, Kim and Poor, 2011). These approaches commonly assume the electric utility has perfect information on the motivational state of each homeowner, and that the uncertainty is primarily in electricity generation and pricing. In contrast, our proposed methodology has the ability to estimate the motivational state of each homeowner to better design DR programs through improved targeting of price-based and appliance-replacement incentives. (Existing work also does not consider the option of the power company to provide rebates for upgrading inefficient equipment, while our framework can incorporate this scheduling problem.)

### Decentralized Supply Chain Logistics

Consider a supply chain in which a single manufacturer is distributing its products through a network of several retailers that are managed independently of the manufacturer, where these retailers can decide on factors such as purchase quantity, inventory, and product shelf placement. The manufacturer can set incentives as part of a contractual agreement with the retailers in order to maximize their profits and ensure a favorable future relationship with the retailers. These incentives can take the form of different types of promotions such as volume based discounts or reimbursements on currently stocked inventory in exchange for better shelf placement and larger wholesale orders from the retailer (Feighery et al., 2003). However, retailers serve heterogeneous markets, and have heterogeneous resources and managerial priorities. As such, the manufacturer should take into consideration various retailer-specific parameters when negotiating contracts to better achieve their goals.

Over the past several decades, many companies have invested in information sharing programs to aide production planning and retail inventory planning in decentralized supply chains (Sahin and Robinson, 2002). In particular, sharing of information between retailers and manufacturers has become easier due to the implementation of modern point of sale systems and ERP software (Lee and Whang, 2000). However, while some information is shared, for various reasons the retailer firms may not want to have perfect information sharing with the manufacturer (Corbett and Tang, 1999). Therefore the manufacturer could potentially improve their contract design with each of the retailers by using the data which is shared with them to estimate various unknown retailer parameters (e.g. local market conditions, internal cost structure, etc.). Since the contracts are often renegotiated on a quarterly or bi-annual basis, using an adaptive method for contract design could improve the construction of future contracts.

There is a significant amount of literature on modeling this relationship between retailers and manufacturers as a repeated Stackelberg game (He et al., 2007). In particular, several different model dynamics have been considered for various contract designs such as seasonal demand with the contract quantity being a fixed wholesale price (Eliashberg and Steinberg, 1987, Desai, 1996) and time varying wholesale cost contracts (Desai, 1992), and demand that depends on the pricing policy of the retailer and contracts that include wholesale price and shelf space allocation (He and Sethi, 2008, Gutierrez and He, 2011, Kogan and Tapiero, 2007). The majority of this work considers models with only a single retail agent, and perfect information sharing between the retailer and manufacturer. In contrast to these assumptions, the behavioral analytics framework we consider in this chapter can be seen as an extension of these models for the case where the manufacturer has imperfect information of retailer specific information (e.g. internal organization, local market conditions etc.) and must estimate these unknown factors. Additionally, much of this existing work is focused on analyzing the equilibrium behavior of these supply chain systems and on characterizing closed form policies for the optimal contracts. The resulting closed form policies are thus heavily dependent on the specific dynamics and contract structures used in each scenario and cannot be easily adapted to changing conditions. In contrast, our behavioral analytics framework for adaptively designing incentives – which consists of repeatedly estimating utility functions and then refining the incentives using optimization modeling – is focused on computing a policy which improves as more data is collected on the behavior of the retailers. Our framework can be used to analyze a more general class of contract structures and retailer behavior with additional operational flexibility at the cost of not having a closed form for the optimal policy.

### 3.1.2 Literature Review

The behavioral analytics framework we develop in this chapter builds upon existing literature on data driven and adaptive methods for stochastic optimization. Ban and Rudin (2016) and Vahn (2015) consider how predictive and data driven models can be incorporated into inventory management problems, and both parametric and nonparametric predictive models are used by a decision maker to estimate demand and compute an optimal reorder policy. These models are constructed to estimate demand through i.i.d observations; this differs from the setup in this chapter where the observations are generated by temporal dynamics and are thus not i.i.d. A more general set of approaches are reinforcement learning and Bayesian optimization (Aswani et al., 2013, Frazier and Wang, 2016, Osband and Roy, 2015, Osband et al., 2016), which leverage statistical estimation to compute asymptotically optimal control inputs for systems with appropriate model structures. However, the relationship between the computed control inputs and the estimated model is often difficult to interpret because of the nonparametric nature of the estimation (Breiman et al., 2001). Our approach offers improved interpretability of the incentives computed by our framework because we simultaneously generate estimates of the parameters of the utility function (i.e., motivational states) for each agent. These estimates provide insights into the resulting incentive allocations computed by our framework because these parameters usually have behavioral or financial interpretations (e.g., responsiveness to incentives, production efficiency, level of risk aversion).

Our behavioral analytics framework is also related to research that explores stochastic control of multi-agent systems. Related methods include decentralized control (Li et al., 2012), approximate dynamic programming (Boukhtouta et al., 2011, George and Powell, 2007), game-theoretic approaches (Adlakha and Johari, 2013, Iyer et al., 2011, 2014, Zhou et al., 2016), and robust optimization (Blanchet et al., 2013, Bertsimas and Goyal, 2012, Lorca and Sun, 2015). In general, these models consider very different settings from the ones we consider in this chapter. This body of work studies settings where the agents can strategically interact with other agents (without the presence of a coordinator) and where the agents are able to consider long time horizons when making decisions. Our setting differs in that we have a single coordinator that provides incentives to a group of agents that do not interact strategically with other agents, and where the agents are myopic (meaning they make decisions based on short time horizons). The three examples of weight loss programs, demand response programs, and franchise logistics more closely match the setting we consider in this chapter.

### 3.1.3 Contributions

Our overall goal in this chapter is to provide tools and approaches that form a specific implementation of the three steps of a behavioral analytics framework, and our secondary goal is to give an example that demonstrates how our implementation of behavioral analytics can be applied to a real-world engineering problem. Recall that these three steps involve designing a behavioral model, and then repeatedly estimating the parameters of this behavioral model, and using the estimated parameters to optimize the incentives provided to each agent. To do this, we first need to identify a general (and practically useful) class of models that describe agent behavior and can be incorporated into optimization models for incentive design. This is non-obvious because incentive design in principle requires solving bilevel programs, precluding the straightforward use of commercial optimization software packages. We address this by abstracting and generalizing our earlier work on the development of predictive models of the behavior of individuals participating in a weight loss program (Aswani et al., 2016). Given these behavioral models, we design an optimization approach that, rather than directly solving the relevant bilevel program, is built around formulations that incorporate the individual behavior model into mathematical programs that can be solved in a straightforward way with commercial solvers, and that lead to incentives that are asymptotically optimal as more data is collected. Below, we describe these contributions in further detail:

First, we develop and analyze an abstract model of agent behavior. This model consists of a myopic utility function (meaning the agent makes decisions based on a utility function that depends on states only one time period into the future) and temporal dynamics on the system states and on the parameters of the utility function. It abstracts and generalizes a predictive model we created in our prior work on behavioral modeling for weight loss (Aswani et al., 2016). In addition, we explore (for the first time) theoretical questions related to statistical consistency of utility function parameter estimates. Such consistency is important because in order to design optimal incentives we need to be able to correctly estimate the parameters of the utility functions of each agent, and it was recently shown that not all approaches that have been proposed for estimating parameters of utility functions are statistically consistent (Aswani et al., 2018). Here, we provide mixed integer linear programming (MILP) formulations for estimating the parameters of the utility functions, and we prove these formulations generate estimates that are statistically consistent.

We also develop novel mathematical programs for incentive design that incorporate our model for agent behavior, and we prove that the incentives are asymptotically optimal (in time). Incentive design in principle requires solving a bilevel program, and the situation is complicated in our setting because the mathematical structure of our abstract model for agent behavior leads to incentive design problems that consist of bilevel mixed integer programs (BMIP's). BMIP's are computationally difficult to solve (Ralphs and Hassanzadeh, 2014, DeNegre and Ralphs, 2009, James and Bard, 1990, Moore and Bard, 1992) since solution techniques for continuous bilevel programming (Ahuja and Orlin, 2001, Aswani et al., 2018, Ouattara and Aswani, 2018, Dempe, 2002, Heuberger, 2004) cannot be used. Consequently, we develop an adaptive two-stage decomposition algorithm. In the first stage, we solve the coordinator's problem for each agent considered individually by estimating the utility function parameters of an agent by solving a single MILP and then solving a series of MILP sub-problems. The second stage consists of an integer linear program (ILP) master problem that aggregates the sub-problem solutions and solves the coordinator's problem for all agents considered jointly. We prove this asymptotically designs the optimal incentives.

To evaluate the efficacy of the specific behavioral models, parameter estimation techniques, and optimization models in our instantiation of a behavioral analytics framework, we perform computational experiments in the context of goal-setting and clinical appointment scheduling for individuals participating in a clinically-supervised weight loss program. The first step of our behavioral analytics approach involves constructing a model that describes individuals' decisions on how much to eat and how much physical activity (in terms of daily steps) to do – subject to a utility function that captures the tradeoffs inherent in achieving one-day-ahead weight loss with reducing dietary consumption and increasing physical activity. The second step of our behavioral analytics approach uses past data for each individual in order to quantify (for each individual) the tradeoffs captured by the utility function, as well as estimate the responsiveness of each individual to the incentives of providing physical activity goals and providing counseling sessions, and the third step of our behavioral analytics approach uses the behavioral model and estimated parameters to determine what physical activity goals to provide to each individual and to determine how to allocate a fixed number of counseling sessions to a pool of individuals participating in the program. These second and third steps are repeated as more data is collected from each individual. Through a simulation study, we compare personalized treatment plans computed by our approach with treatment plans computed by an adaptive heuristic, and we find that our approach performs substantially better than the heuristic. Common heuristics implicitly assume monotonicity in individuals' behaviors with respect to the treatment plan values, while actual behavior (captured by our predictive models) displays substantial non-monotonicity: For example, losing weight causes individuals to eat more and exercise less, so the speed of weight loss can impact the final weight loss outcomes.

### 3.1.4 Outline

Section 3.2 describes the first step of our behavioral analytics framework – the development of the behavioral model. The model consists of a utility function – describing how an agent makes decisions – and temporal dynamics on the system states and parameters of the utility function. We refer jointly to both components of this abstract model as the behavioral model. Section 3.3 presents approaches for estimating parameters of this behavioral model using MILP formulations to solve the problems of maximum likelihood estimation (MLE) and Bayesian inference. We prove that solutions of our MILP formulations provide consistent estimates of the agent's parameters. In Section 3.4, we present algorithms for optimizing the incentives provided to agents by the coordinator. We first present an algorithm based on solving two MILP's that allows the coordinator to allocate incentives in the situation where there is only a single agent with unknown-to-the-coordinator parameters, and we prove that this algorithm computes incentives that are asymptotically optimal (in the sense of minimizing the coordinator's loss function) as time $t$ goes to infinity. Next, we develop a two-stage decomposition algorithm (building on the single-agent formulation) to solve the coordinator's problem in a multi-agent setting, and we generalize our proof of asymptotic optimality to this setting Finally in Section 3.5, we study (via simulation) the effectiveness of our algorithms for designing personalized weight loss treatment plans. Our results show that treatment plans computed by our behavioral analytics approach could potentially reduce the cost of running such weight loss programs by as much as 60% without affecting the efficacy of these programs.

## 3.2 Predictive Modeling of a Single Myopic Agent

In this section, we present our behavioral model for a single myopic agent. This forms the first step of our specific implementation of a behavioral analytics framework, and the key design problem is formulating a predictive model that is amenable to performing the second and third steps of our behavioral analytics framework of parameter estimation and incentive optimization. This model is an extension and abstraction of a behavioral model that was validated in our past work on behavioral modeling for weight loss (Aswani et al., 2016), in which we used cross-validation (i.e., out-of-sample comparisons) to perform a data-based validation of the predictive accuracy of our behavioral model by comparison to a standard machine learning algorithm for prediction.

Let $\mathcal{X}, \mathcal{U}, \Pi, \Theta$ be compact finite-dimensional sets with $\mathcal{X}, \mathcal{U}, \Theta$ convex. We will refer to the agent's system states $x_t \in \mathcal{X}$, motivational states (or type) $\theta_t \in \Theta$, and decisions $u_t \in \mathcal{U}$ at time $t$. The coordinator provides an incentive (or input) $\pi_t \in \Pi$ to the agent at time $t$, and we assume that the motivational states are unknown to the coordinator but known to the agent. In our behavioral model, the system and motivational states are subject to temporal dynamics:

$$\begin{aligned} x_{t+1} &= h(x_t, u_t), \\ \theta_{t+1} &= g(x_t, u_t, \theta_t, \pi_t). \end{aligned} \tag{3.1}$$

The intuition of the above dynamics is that future system states $x_{t+1}$ depend on the current system states $x_t$ and decision $u_t$, while future motivational states $\theta_{t+1}$ depend on the current system states $x_t$, decision $u_t$, motivational states $\theta_t$, and incentives $\pi_t$.

The agents are modeled to be myopic in the sense that agents make decisions at time $t$ by considering only their present utility function. We assume the agent's utility function belongs to a parametrized class of functions $\mathcal{F} := \{(x, u) \mapsto f(x, u, \theta, \pi) : \theta \in \Theta, \pi \in \Pi\}$; and the agent's utility function at time $t$ is $f(\cdot, \cdot, \theta_t, \pi_t)$. Thus at time $t$ the agent's decisions are

$$u_t \in \operatorname{argmax} \{f(x_{t+1}, u, \theta_t, \pi_t) \mid x_{t+1} = h(x_t, u),\ u \in \mathcal{U}\}, \tag{3.2}$$

which means we are assuming the agent has perfect knowledge of $x_t, \theta_t, \pi_t$. This model says that the agent's decisions depend on the current system states, motivational states, and incentives. For notational simplicity, we have assumed each agent has the same $f$, $h$, $g$; however, our behavioral analytics framework immediately generalizes to a setting where these functions are different for each agent. To reflect this in terms of notation, we would replace these functions with the functions $f^a$, $h^a$, $g^a$ for $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of agents.

Though the coordinator also has perfect knowledge of the incentives $\pi_t$, the coordinator can only make noisy observations of past system states and agent decisions:

$$\begin{aligned} \tilde{x}_{t_i} &= Dx_{t_i} + \nu_{t_i} \quad \forall i = 0, \ldots, n_x, \\ \tilde{y}_{\tau_i} &= Cu_{\tau_i} + \omega_{\tau_i} \quad \forall i = 0, \ldots, n_u, \end{aligned} \tag{3.3}$$

where $C, D$ are known output matrices, and $x_{t_i}, u_{\tau_i}$ are the systems states and agent decisions generated by (3.1) and (3.2) with initial conditions $(x_0, \theta_0)$ and incentives $(\pi_1, \pi_2, \ldots)$. Here, the sequences $\{t_i\}_{i=0}^{n_x}$ and $\{\tau_i\}_{i=0}^{n_u}$ denote the time instances at which noisy measurements of the system state and agent's decisions are made, respectively. Similarly, $n_x$ and $n_u$ are the number of measurements of the system state and agent's decisions that have been made, respectively. Observe that in our setting the coordinator has complete information about the temporal dynamics of system and motivational states, while the state observations are noisy. This is a realistic assumption for many applications, such as the ones described in the introduction. For instance, we have recently completed a rigorous clinical trial to experimentally validate the efficacy of our behavioral analytics framework in a setting where the aim of the coordinator was to increase the physical activity of individuals through a mobile phone app (Zhou et al., 2018). However, some settings may involve scenarios where the temporal dynamics of system and motivational states are unknown to the coordinator, and our approach described in this chapter will not be applicable to this more challenging scenario. If the dynamics are unknown, the coordinator would be required to use an approach such as reinforcement learning (Aswani et al., 2013, Frazier and Wang, 2016, Osband and Roy, 2015, Osband et al., 2016), but the disadvantage of reinforcement learning is that its convergence rate in time to the optimal policy will be slower because of the need to estimate dynamics. Interpreted in this way, our work develops an approach that lies between deterministic optimization with complete information and reinforcement learning in terms of the information known about the overall system.

For our subsequent optimization modeling and theoretical analysis, we make the following assumptions about this behavioral model:

**Assumption 3.1.** The sets $\mathcal{X}, \mathcal{U}, \Pi, \Theta$ are bounded and finite-dimensional. Moreover, the sets $\mathcal{X}, \mathcal{U}, \Theta$ are convex polyhedra described by a finite number of linear inequalities, and $\Pi$ can be described by a finite number of mixed integer linear constraints.

This mild assumption ensures that states, decisions, and inputs are bounded; that the range of possible values for states and inputs are polytopes; and that the set of possible incentives is representable by mixed integer linear constraints.

**Assumption 3.2.** The function $f : \mathcal{X} \times \mathcal{U} \times \Theta \times \Pi \to \mathbb{R}$ is deterministic, concave in $x$, strictly concave in $u$, and concave in $\theta$; moreover, $f$ can be expressed as

$$f(x, u, \theta, \pi) = -(x; u)^T \cdot Q \cdot (x; u) + (\theta; \pi)^T \cdot H \cdot (x; u) + \sum_{i=1}^{K} \min_{j \in J_i} \{F_{i,j} \cdot (x; u; \theta; \pi) + \zeta_{i,j}\}, \quad (3.4)$$

where $Q$ is a positive semidefinite matrix, the $F_{i,j}, H$ are matrices of appropriate dimension, the $\zeta_{i,j}$ are scalars, $K$ is a positive scalar corresponding to the number of piecewise linear components, and the $J_i$ are sets of indices where each index corresponds to a particular linear function which forms the piecewise linear component $i$.

Strict concavity in $u$ ensures $\operatorname{argmax}_{u \in \mathcal{U}} f(x, u, \theta, \pi)$ is singleton for all $(x, \theta, \pi) \in \mathcal{X} \times \Theta \times \Pi$, and the concavity assumptions also model diminishing returns and ensure $u_t$ is polynomial-time computable by the agent (Brock and Wartman, 1990, Gafni, 1990, Cawley, 2004). The specific form of $f$, a combination of concave quadratic and piecewise linear terms, is useful for modeling as it allows for a rich family of functions that in practice can be used to approximate various possible utility functions, while also ensuring that the states can be estimated using a MILP.

**Assumption 3.3.** The functions $h : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ and $g : \mathcal{X} \times \mathcal{U} \times \Theta \times \Pi \to \Theta$ are deterministic surjective functions of the form

$$\begin{aligned} &h(x, u) = Ax + Bu + k \\ &g(x, u, \theta, \pi) = G_i \cdot (x; u; \theta; \pi) + \xi_i \text{ when } B_i \cdot (x; u; \theta; \pi) \leq \psi_i \end{aligned} \tag{3.5}$$

where $A, B, G_i, B_i$ are matrices; $\gamma_i, \psi_i, k$ are vectors; $\xi_i$ are scalars; and the interiors of the polytopes $B_i \cdot (x; u; \theta; \pi) \leq \psi_i$ are disjoint.

This condition on $h, g$ allows us to formulate problems of statistical estimation as a MILP. The specific form of $h$ (i.e., a linear function) and $g$ (i.e., a piecewise linear function) is useful for modeling as it allows for a rich family of functions that in practice can be used to approximate various possible dynamics, while also ensuring that the states can be estimated using a MILP.

**Assumption 3.4.** The $\{\nu_{t_i}\}_{i=0}^{n_x}$ and $\{\omega_{\tau_i}\}_{i=0}^{n_u}$ from the measurement noise model (3.3) are sequences of i.i.d random vectors with i.i.d components with zero mean and (known) finite variance. Moreover, the logarithm of their probability density functions can be expressed using integer linear constraints and integer linear objective terms.

This means $\mathbb{E}\omega_{\tau_i} = \mathbb{E}\nu_{t_i} = 0$ and $\mathbb{E}(\nu_{t_i})_j^2 = \sigma_\nu^2 < \infty$ and $\mathbb{E}(\omega_{\tau_i})_j^2 = \sigma_\omega^2 < \infty$ with known $\sigma_\nu^2, \sigma_\omega^2$. Furthermore, the density functions can be reformulated so that the estimation problem is amenable to MILP solvers. Examples of noise distributions satisfying the integer linear representability assumption include the Laplace distribution, the shifted exponential distribution, and piecewise linear distributions. This assumption can be relaxed to requiring integer quadratic representability (such as is the case for Gaussian distributed noise), and the subsequent results change in that the optimization formulations become MIQP's, rather than the MILP's that occur with the above assumption.

**Assumption 3.5.** The discrete-time system with temporal dynamics (3.1) and (3.2) and measurement model (3.3) is observable (i.e., there exists a $T$ and sequence $\pi_t$ such that $(x_0, \theta_0)$ can be exactly computed if the measurements from $0 \leq t \leq T$ are noiseless).

This last assumption is an identifiability condition (Bickel and Doksum, 2006), meaning that different initial conditions $(x_0, \theta_0)$ on the agent's system and motivational states produce different sequences of measurements and states, and this is a common assumption for control systems (Callier and Desoer, 1994). This assumption is equivalent to assuming that exact knowledge of the current state completely characterizes all past and future states (Callier and Desoer, 1994), and it is a frequently-made assumption because it ensures enough predictability in the state trajectories so that theoretical results can be proved. However, in some real applications it may be that as time proceeds, the impact of the initial condition of the state may become more negligible. In such cases, a common heuristic known as a “forgetting factor” (Fortescue et al., 1981, Dasgupta and Huang, 1987, Nelles, 2001, Leung and So, 2005) is often used. This heuristic modifies the optimization problem that is solved to perform state estimation by placing an exponentially-decaying-in-time weight on older data. It is in this way that the “forgetting factor” heuristic emphasizes more recent measurements and allows the estimation to become less sensitive to the initial condition. Though theoretical results can sometimes be proved for frameworks with a “forgetting factor”, we do not consider this extension here because it requires substantial additional analysis that is beyond the scope of this chapter. The second assumption is common for utility functions (Brock and Wartman, 1990, Gafni, 1990, Cawley, 2004). The third assumption says the system state dynamics are linear, and that the motivational state dynamics are piecewise affine, which are common models for control systems (Callier and Desoer, 1994, Mignone et al., 2000, Aswani and Tomlin, 2009). We believe all five assumptions are satisfied by agents in the three examples of weight loss programs, demand response programs, and decentralized supply chain management. Section 3.5 provides a behavioral model for agents in a weight loss program that satisfies our above assumptions, and we conclude that section with a computational study where we solve the coordinator's problem for a weight loss program.

## 3.3 Estimating Model Parameters

In this section, we explore how the coordinator can estimate the agent's initial states $(x_0, \theta_0)$, and predict the agent's future behavior for a fixed policy $\pi$. We refer to these states $(x_0, \theta_0)$ as the agent model parameters (not to be confused with the structural parameters of the model: $f, g$, and $h$) since by Assumptions 3.1–3.5 these values completely characterize the behavior of the system and are not fully known to the coordinator. This forms the second step of our implementation of a behavioral analytics framework, and we leverage the mathematical structure of the behavioral model described in Section 3.2 to construct techniques and methods for estimation and prediction. This second step is important because the estimated parameters of the behavioral model and subsequent predictions of future agent behavior are used to optimize incentives in the third step of our behavioral analytics approach. We will assume the coordinator makes noisy and partial observations – according to the measurement model (3.3) – of the agent's state and decisions for $n$ time periods (with some missing observations). In Section 3.3.1, we present a Maximum Likelihood Estimation approach to estimate the agent's initial system states and motivational states. In Section 3.3.2, we consider a setting in which the coordinator has some prior knowledge about the possible values of the motivational states, and consider a Bayesian setting.

### 3.3.1 Maximum Likelihood Estimation

Let $\{\tilde{x}_{t_i}\}_{i=0}^{n_x}$ denote the process of the state observations, and let $\{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}$ denote the process of the behavior observations.

Our approach to estimating the agent's initial parameters will be to compute estimates $(\hat{x}_0, \hat{\theta}_0) \in \text{argmin}_{(x_0,\theta_0)\in\mathcal{X}\times\Theta}\, \mathcal{L}(x_0, \theta_0, \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}; \pi)$ by minimizing an appropriately chosen loss function $\mathcal{L}$. More specifically, we use the approach of Maximum Likelihood Estimation (MLE), which is equivalent to choosing a loss function that corresponds to the negative likelihood. Let $p_\nu, p_\omega$ be the density functions of $\nu_{t_i}, \omega_{\tau_i}$; then the joint likelihood function of $(\theta_0, x_0)$ for a fixed $\pi$ is

$$\begin{aligned}\mathcal{L}(x_0, \theta_0, \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \pi) &= p(\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u} | \theta_0, x_0, \pi) \\ &= \prod_{i=0}^{n_x} p_\nu(\tilde{x}_{t_i} - Dx_{t_i}) \prod_{j=0}^{n_u} p_\omega(\tilde{y}_{\tau_j} - Cu_{\tau_j})\end{aligned} \tag{3.6}$$

Thus the coordinator's estimation problem is given by the following:

$$\begin{aligned}(\hat{x}_0, \hat{\theta}_0) \in \underset{\{(x_t,\theta_t,u_t)\}_{t=0}^{T}}{\text{argmax}} & \sum_{i=0}^{n_x} \log p_\nu(\tilde{x}_{t_i} - Dx_{t_i}) + \sum_{j=0}^{n_u} \log p_\omega(\tilde{y}_{\tau_j} - Cu_{\tau_j}) \\ \text{s.t.} \quad & \begin{aligned} u_t \in \text{argmax} & f(x_{t+1}, u, \theta_t, \pi_t) \\ \text{s.t. } & x_{t+1} = h(x_t, u), u \in \mathcal{U} \end{aligned} && 0 \le t \le T-1, \\ & \theta_{t+1} = g(x_t, u_t, \theta_t, \pi_t) && 0 \le t \le T-1, \\ & x_t \in \mathcal{X}, \theta_t \in \Theta && 0 \le t \le T.\end{aligned} \tag{3.7}$$

Problem (3.7) is a bilevel optimization problem because the $u_t$ are minimizers of $f(x_{t+1}, \cdot, \theta_t, \pi_t)$, and such bilevel problems frequently arise in the context of estimating utility functions (Keshavarz et al., 2011, Bertsimas et al., 2014, Aswani et al., 2018). We note that the constraints are in effect for all $t$ and not simply for the values where observations are collected, which ensures that the optimization problem can account for missing observations by imputing parameter values using the model dynamics. For the setting we consider in this chapter, we show that the bilevel program for MLE (3.7) can be exactly reformulated as a MILP.

**Proposition 3.1.** If Assumptions 3.1–3.5 hold; then the feasible region of (3.7) can be formulated as a set of mixed integer linear constraints with respect to $(x_t, u_t, \theta_t, \pi_t)$.

The full proof for this proposition can be found in Appendix B.1 but here we will provide some intuition for this proof. First we note that by Assumption 3.3 the constraints $\theta_{t+1} = g(x_t, u_t, \theta_t, \pi_t)$ can be linearized using a big $M$ formulation and the additions of binary variables $\iota_i$ that indicate whether the parameters are contained in polytopes $B_i \cdot (x; u; \theta; \pi) \leq \psi_i$. For the remainder of the proof, we show that the optimality set of $f(x, u, \theta, \pi)$ can be expressed as a set of mixed integer linear constraints. Using assumptions 3.2 and 3.3, we can reformulate this optimization problem as a constrained convex quadratic program with a strictly concave objective in $u_t$. This means that the first order KKT conditions are both necessary and sufficient to describe $u_t$ as the maximizer of $f$, and these conditions can be written in a mixed integer linear form.

An important consequence of this proposition is that it is possible to compute the global solution of the MLE problem (3.7) using standard optimization software.

**Corollary 3.2.** If Assumptions 3.1–3.5 hold, then the MLE problem (3.7) can be expressed as a MILP.

**Remark 3.1.** If the logarithm of the noise densities can be expressed using integer quadratic constraints (e.g., Gaussian distributions), then the MLE problem (3.7) can be expressed as a MIQP.

### 3.3.2 Bayesian Estimation

Solving the MLE problem (3.7) gives an estimate of the agent's initial system states and motivational states, which completely characterize the agent. However, the coordinator often has some prior knowledge about the possible values of the motivational states. In such a case, a Bayesian framework is a natural setting for making predictions of the agent's future system states.

Suppose the coordinator has interacted with the agent over $T$ time periods, has measured $\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}$ with $T_{n_x}, T_{n_u} \leq T$, and wants to predict the agent's future states and decisions $\{x_i, \theta_i, u_i\}_{i=T}^{T+n}$ for some $n > 0$ time steps into the future. In principle, this means the coordinator wants to calculate the posterior distribution of $\{x_i, \theta_i, u_i\}_{i=T}^{T+n}$. But $(x_0, \theta_0)$ completely characterize the agent in our model (recall Assumption 3.5 states that distinct initial conditions produce different state and decision trajectories), and so we can predict the agent's future states and decisions using the posterior distribution of $(x_0, \theta_0)$. Hence we focus on computing the posterior of $(x_0, \theta_0)$. A direct application of Bayes's Theorem (Bickel and Doksum, 2006) gives

$$p(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T+n}) = \\ Z^{-1} \times p(\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u} | x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n}) \times p(x_0, \theta_0). \quad (3.8)$$

Here $Z$ is a normalization constant that ensures the right hand side is a probability distribution, and $p(x_0, \theta_0)$ reflects the coordinator's prior beliefs. We begin with an assumption on $p(x_0, \theta_0)$.

**Assumption 3.6.** The function $\log p(x_0, \theta_0)$ can be expressed using a finite number of mixed integer linear constraints, and $p(x_0, \theta_0) > 0$ for all $(x_0, \theta_0) \in \mathcal{X} \times \Theta$.

This is a mild assumption because it holds for the Laplace distribution, the shifted exponential distribution, and piecewise linear distributions. Significantly, it is true when the prior distribution $p(x_0, \theta_0)$ is an empirical histogram with data in each histogram bin (Aswani et al., 2016).

Next, we describe an optimization approach to computing the posterior distribution of $(x_0, \theta_0)$. Consider the following feasibility problem for fixed initial conditions $(\bar{x}_0, \bar{\theta}_0)$:

$$\begin{aligned}
\psi_T(\bar{x}_0, \bar{\theta}_0) &= \log p(x_0 = \bar{x}_0, \theta_0 = \bar{\theta}_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T+n}) + \log Z \\
&= \max_{\{(x_t, \theta_t, u_t)\}_{t=0}^{T}} \sum_{i=0}^{n_x} \log p_\nu(\tilde{x}_{t_i} - D x_{t_i}) + \sum_{j=0}^{n_u} \log p_\omega(\tilde{y}_{\tau_j} - C u_{\tau_j}) + \log p(x_0, \theta_0) \\
&\qquad \text{s.t.} \quad \begin{array}{ll}
\begin{array}{l} u_t \in \operatorname{argmax} f(x_{t+1}, u, \theta_t, \pi_t) \\ \qquad \text{s.t. } x_{t+1} = h(x_t, u), u \in \mathcal{U} \end{array} & 0 \le t \le T-1, \\
\theta_{t+1} = g(x_t, u_t, \theta_t, \pi_t) & 0 \le t \le T+n-1, \\
x_0 = \bar{x}_0, \theta_0 = \bar{\theta}_0, & \\
x_t \in \mathcal{X}, \theta_t \in \Theta & 0 \le t \le T+n.
\end{array}
\end{aligned} \tag{3.9}$$

The above problem is almost the same as the MLE problem (3.7), with the only differences that the above has additional constraints $x_0 = \bar{x}_0, \theta_0 = \bar{\theta}_0$ and an additional term in the objective $\log p(x_0, \theta_0)$. Thus we have that the above problem (3.9) can be expressed as a MILP or MIQP.

**Corollary 3.3.** If Assumptions 3.1–3.6 hold, then (3.9) can be formulated as a MILP.

**Remark 3.2.** Under appropriate relaxed representability conditions on the noise distributions and the prior distribution, the problem (3.9) can be formulated as a MIQP.

Solving (3.9) does not directly provide the posterior distribution of $(x_0, \theta_0)$ because $Z$ is not known *a priori*, though it can be computed using numerical integration. (See for instance the approach by Aswani et al. (2016).) But since $Z$ only scales the posterior estimate, we instead propose a simpler scaling. Let $(\hat{x}_{0,T}, \hat{\theta}_{0,T}) \in \operatorname{argmax}_{(x_0, \theta_0)} \psi_T(x_0, \theta_0)$ be the maximum *a posteriori* (MAP) estimates of the initial conditions, and note that the above corollaries apply to the computation of the MAP because the corresponding optimization problem for computing the MAP is simply (3.9) but with the constraints $x_0 = \bar{x}_0, \theta_0 = \bar{\theta}_0$ removed. We propose using

$$\hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) = \frac{\exp(\psi_T(x_0, \theta_0))}{\exp(\psi_T(\hat{x}_{0,T}, \hat{\theta}_{0,T}))} \tag{3.10}$$

as an estimate of the posterior distribution of $(x_0, \theta_0)$. Two useful properties of our estimate are that $\hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \in [0,1]$ by construction, and that $\hat{p}(\hat{x}_{0,T}, \hat{\theta}_{0,T} | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) = 1$ by construction. We will show this estimate is statistically consistent in a Bayesian sense (Bickel and Doksum, 2006):

**Definition** The posterior estimate (3.10) is consistent if for all $(x_0^*, \theta_0^*) \in \mathcal{X} \times \Theta$ and $\epsilon, \delta > 0$ we have $p_{(x_0^*, \theta_0^*)}(\hat{p}(\mathcal{E}(\delta) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \geq \epsilon) \to 0$ as $T \to \infty$, where $p_{(x_0^*, \theta_0^*)}$ is the probability law under $(x_0^*, \theta_0^*)$, $\mathcal{E}(\delta) = \{(x_0, \theta_0) \notin \mathcal{B}(x_0^*, \theta_0^*, \delta)\}$, and $\mathcal{B}(x_0^*, \theta_0^*, \delta)$ is an open $\delta$ ball around $(x_0^*, \theta_0^*)$.

The meaning of this definition is that if $(x_0^*, \theta_0^*)$ are the true initial conditions of the agent, then a consistent posterior estimate is such that it collapses until all probability mass is on the true initial conditions. Statistical consistency of (3.10) also needs an additional technical assumption:

**Assumption 3.7.** Let $(x_0^*, \theta_0^*)$ be the agent's true initial conditions. The incentives $\pi_t$ are such that

$$\max_{\mathcal{E}(\delta)} \lim_{T \to \infty} \sum_{i=0}^{n_x} \log \frac{p_\nu(\tilde{x}_{t_i} - D\overline{x}_{t_i})}{p_\nu(\tilde{x}_{t_i} - Dx_{t_i})} + \sum_{j=0}^{n_u} \log \frac{p_\omega(\tilde{y}_{\tau_j} - C\overline{u}_{\tau_j})}{p_\omega(\tilde{y}_{\tau_j} - Cu_{\tau_j})} = -\infty \tag{3.11}$$

for any $\delta > 0$, almost surely, where $x_t, u_t$ are the states and decisions under initial conditions $(x_0^*, \theta_0^*)$, and $\overline{x}_t, \overline{u}_t$ are the states and decisions under initial conditions $(x_0, \theta_0)$.

This type of assumption is common in the adaptive control literature (Craig et al., 1987, Astrom and Wittenmark, 1995), and is known as a *sufficient excitation* or a *sufficient richness* condition. It is a mild condition because there are multiple ways of ensuring this condition holds (Bitmead, 1984, Craig et al., 1987, Astrom and Wittenmark, 1995). One simple approach (Bitmead, 1984) is to compute an input $\pi_t$ and then add a small amount of random noise (whose value is known since it is generated by the coordinator) to the input before applying the input to the agent.

**Proposition 3.4.** If Assumptions 3.1–3.7 hold, then the estimated posterior distribution denoted by $\hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T})$ and given in (3.10) is consistent.

The full proof of this proposition can be found in Appendix B.1 but here we will provide some intuition for the proof. First suppose that the true initial conditions of the system $(x_0^*, \theta_0^*)$ are known. If this is the case then the log of the posterior likelihood of having any other initial conditions $(x_0, \theta_0)$ can be expressed as the posterior likelihood of the initial conditions minus the log likelihood ratio between the distribution generated by $(x_0^*, \theta_0^*)$ and $(x_0, \theta_0)$. Since the ratio of the prior distributions is constant, and the posterior likelihood of $(x_0^*, \theta_0^*)$ is between 0 and 1, we observe that these terms are negative. This means the remaining terms follow the form of Assumption 3.6, and by this assumption this implies that for any $\delta > 0$ the posterior likelihood of any $(x_0, \theta_0)$ that are not within a $\delta$ ball of $(x_0^*, \theta_0^*)$ approaches zero. Hence it follows that this posterior likelihood estimate is consistent.

**Corollary 3.5.** If Assumptions 3.1–3.7 hold, then $(\hat{x}_{0,T}, \hat{\theta}_{0,T}) \xrightarrow{p} (x_0^*, \theta_0^*)$ as $T \to \infty$.

A full proof of this corollary can be found in Appendix B.1 but here we will provide some intuition for this proof. First we consider the event that the MAP estimator $(\hat{x}_{0,T}, \hat{\theta}_{0,T})$ is not within a $\delta$ ball around $(x_0^*, \theta_0^*)$ for some $\delta > 0$. This event is contained in the event that the largest value of the posterior likelihood outside this $\delta$ ball is greater then the largest value of the posterior likelihood inside the ball. However, by Proposition 3.4 we see that the posterior probability measure must concentrate about $(x_0^*, \theta_0^*)$ for the probability law $p_{(x_0^*, \theta_0^*)}$. This means that as $T \to \infty$, the probability of the maximum being outside $\mathcal{B}(x_0^*, \theta_0^*, \delta)$ approaches zero thus completing the proof.

The above two results imply that future agent behavior can be reasonably predicted using the MAP parameters. Recall that calculating the MAP can be formulated as a MILP or MIQP, since the corresponding optimization problem is (3.9) with the constraints $x_0 = \bar{x}_0, \theta_0 = \bar{\theta}_0$ removed.

## 3.4 Optimizing Incentives

The final step of our behavioral analytics framework involves using estimates of behavioral model parameters for each agent to optimize the design of costly incentives provided to the agents by the coordinator. In Section 3.4.1, we develop an algorithm for the single agent case. In Section 3.4.2, we use this single-agent algorithm as a sub-problem in the multi-agent case. In both cases, we show that our algorithms are asymptotically optimal (as time continues and more data is collected) with respect to the coordinator's loss function when the agents behave according to the model constructed in Section 3.2. The two algorithms we present in fact combine the second and third steps of our framework by first applying the parameter estimation algorithms (described in Section 3.3) that comprise the second step, and then optimizing incentives. The benefit of combining the second and third steps into a single algorithm is that this makes it easier to recompute the incentives as more data is collected over time from each agent.

### 3.4.1 Optimizing Incentives for a Single Agent

Consider the problem of designing optimal incentives for a single agent at time $T$ by choosing $\{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n$ to minimize a bounded loss function $\ell : \mathcal{X}^n \times \mathcal{U}^n \to \mathbb{R}$ of the agent's system states and decisions over the next $n$ time periods. In this and subsequent sections, we use the notation $\ell$ instead of $\mathcal{L}$ for the loss functions in order to signify that we are interested in a function of the decision maker's policy and not an estimation loss. We consider losses of a fairly general form:

**Assumption 3.8.** The loss function $\ell$ can be described by mixed integer linear constraints and mixed integer linear objective terms.

As in the previous section, this assumption assures that $\ell$ can be expressed in a form that can be used with a MILP solver. Since the coordinator only has noisy and incomplete observations of the agent's system states and decisions $\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}$, one design approach is to minimize the expected posterior loss

$$\min \Big\{ \mathbb{E}[\ell(\{x_t, u_t\}_{t=T+1}^{T+n}) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}] \;\Big|\; \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n \Big\}. \tag{3.12}$$

However, recalling our previous discussion, the agent's behavior is completely characterized by the initial conditions $(x_0, \theta_0)$, and so by the sufficiency and the smoothing theorem (Bickel and Doksum, 2006), there exists $\varphi : \mathcal{X} \times \Theta \times \Pi^n \mapsto \mathbb{R}$ such that the design problem can be exactly reformulated as

$$\min \Big\{ \mathbb{E}[\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n}) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}] \;\Big|\; \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n \Big\}. \tag{3.13}$$

Calculating this expectation is difficult because the posterior distribution of $(x_0, \theta_0)$ does not generally have a closed form expression. In principle, discretization approaches from scenario generation (Kaut and Wallace, 2003) could be used to approximate the design problem as

$$\min \Big\{ \frac{\sum_{i=1}^{M} \varphi(x_{i,0}, \theta_{i,0}, \pi) \exp(\psi_T(x_{i,0}, \theta_{i,0}; \pi))}{\sum_{i=1}^{M} \exp(\psi_T(x_{i,0}, \theta_{i,0}; \pi))} \;\Big|\; \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n \Big\}. \tag{3.14}$$

where $(x_{i,0}, \theta_{i,0})$ is an exhaustive enumeration of $\mathcal{X} \times \Theta$. This approximation (3.14) is still challenging to solve because the objective has a fractional, nonconvex form, and $\psi_T$ is defined as the value function of a MILP, meaning that it does not have an easily computable closed form expression (Ralphs and Hassanzadeh, 2014). This means (3.14) is a Bi-level Mixed Integer Program (BMIP) with lower level problems that are MILP's. This is a complex class of optimization problems for which existing algorithms can only solve small problem instances (James and Bard, 1990, Moore and Bard, 1992, DeNegre and Ralphs, 2009).

In this section, we develop a practical algorithm for optimizing incentives for a single agent. We first summarize our algorithm, and show it only requires solving two MILP's. Next we prove this algorithm can be interpreted as solving an approximation of solving either (3.12) or the optimal incentive design problem under perfect noiseless information. More substantially, we also show that our algorithm provides a set of incentives that are asymptotically optimal as time advances.

### Two Stage Adaptive Algorithm (2SSA)

Algorithm 1 summarizes our two stage adaptive approach (2SSA) for designing optimal incentives for a single agent. The idea of the algorithm is to first compute a MAP estimate of the agent's initial conditions, use the MAP estimate as data for the first two arguments of $\varphi$, and then minimize $\varphi$. In fact, we can solve this minimization problem without having to explicitly compute $\varphi$. Because $\varphi$ is defined as the composition of the agent's dynamics with initial conditions $(x_0, \theta_0)$ and the coordinator's loss function $\ell$, it can be written as the value function of a feasibility problem:

$$\begin{aligned}
&\varphi(\overline{x}_0, \overline{\theta}_0, \{\overline{\pi}_i\}_{i=0}^{T+n}) = \\
&\qquad \min_{\{(x_t,\theta_t,u_t,\pi_t)\}_{t=0}^{T+n}} \ell(\{x_t, u_t\}_{t=T+1}^{T+n}) \\
&\qquad\qquad \text{s.t.} \quad \begin{array}{ll}
u_t \in \operatorname{argmax} f(x_{t+1}, u, \theta_t, \pi_t) & \\
\qquad\quad \text{s.t. } x_{t+1} = h(x_t, u), u \in \mathcal{U} & 0 \le t \le T+n-1, \\
\theta_{t+1} = g(x_t, u_t, \theta_t, \pi_t) & 0 \le t \le T+n-1, \\
x_t \in \mathcal{X}, \theta_t \in \Theta, \pi_t \in \Pi & 0 \le t \le T+n, \\
x_0 = \overline{x}_0,\ \theta_0 = \overline{\theta}_0, \pi_t = \overline{\pi}_t & 0 \le t \le T.
\end{array}
\end{aligned} \tag{3.15}$$

More importantly, the problem of minimizing this $\varphi$ can be formulated as a MILP.

**Corollary 3.6.** If Assumptions 3.1–3.8 hold, then $\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n})$ is lower semicontinuous in $x_0, \theta_0, \{\pi_i\}_{i=T+1}^{T+n}$, and the optimization problem given by $\min\{\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$ can be formulated as a MILP for all fixed values of $(x_0, \theta_0, \{\pi_i\}_{i=0}^{T}) \in \mathcal{X} \times \Theta \times \Pi^{T+1}$.

The full proof of this corollary can be found in Appendix B.1 but here we will provide some intuition for the proof of the proposition. First we note that by Proposition 3.1 the feasible region of (3.15) can be expressed as a set of mixed integer linear constraints. Hence $\varphi(\overline{x}_0, \overline{\theta}_0, \{\overline{\pi}_i\}_{i=0}^{T+n})$ is the value function of a MILP in which $x_0, \theta_0, \overline{\pi}_t$ belong to an affine term, and is thus lower semicontinuous. The second result of the proposition follows by noting that the desired optimization problem is equivalent to (3.15) with the removal of the constraints $\pi_t = \overline{\pi}_t$ for $t = T+1, \dots, T+n$. Hence the result follows by the assumptions and Proposition 3.1.

### Asymptotic Optimality of 2SSA

The next result provides the underlying intuition of 2SSA. In particular, we are approximating $\mathbb{E}[\varphi(x_0, \theta_0, \{\pi_i\}_{t=0}^{T+n}) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}]$ using $\varphi(\hat{x}_{0,T}, \hat{\theta}_{0,T}, \{\pi_i\}_{i=0}^{T+n})$, and both these functions are converging to $\varphi(x_0^*, \theta_0^*, \{\pi_i\}_{t=0}^{T+n})$.

**Algorithm 1** Two Stage Single Agent Algorithm (2SSA)

```
Require: $\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{u}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}$
compute $(\hat{x}_{0,T}, \hat{\theta}_{0,T}) \in \arg\max_{(x_0,\theta_0)} \psi_T(x_0, \theta_0)$
return $\pi_{2SSA}(T) \in \arg\min\{\varphi(\hat{x}_{0,T}, \hat{\theta}_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$
```

**Proposition 3.7.** Suppose that Assumptions 3.1–3.8 hold. Then as $T \to \infty$ we have that: $\mathbb{E}[\varphi(x_0, \theta_0, \{\pi_i\}_{t=0}^{T+n}) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}] \xrightarrow{p} \varphi(x_0^*, \theta_0^*, \{\pi_i\}_{t=0}^{T+n})$ for all fixed $\{\pi_i\}_{i=0}^{T+n}$; and $\varphi(\hat{x}_{0,T}, \hat{\theta}_{0,T}, \{\pi_i\}_{t=0}^{T+n}) \xrightarrow[\Pi^n]{\text{l-prob}} \varphi(x_0^*, \theta_0^*, \{\pi_i\}_{t=0}^{T+n})$. Here, $\Lambda_n \xrightarrow[\mathcal{X}]{\text{l-prob}} \Lambda$ means random function $\Lambda_n : \mathcal{X} \to \mathbb{R}$ is a lower semicontinuous approximation to function $\Lambda : \mathcal{X} \to \mathbb{R}$ (Vogel and Lachout, 2003a).

We provide a full proof of this proposition in Appendix B.1 but here we will provide some intuition for the proof. For the first result, we note that since the posterior distribution $p(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T})$ is consistent this means that in the limit it becomes a degenerate distribution centered at $(x_0^*, \theta_0^*)$. Then the first result follows by applying this fact in combination with the dominated convergence theorem. The second result follows by an application of Corollary 3.5 and Corollary 3.5. Essentially, since the MAP estimates are consistent and $\phi$ is lower semicontinuous, then the result follows by using Proposition 2.1.ii of (Vogel and Lachout, 2003b).

If the coordinator had perfect knowledge of the agent's true initial conditions $(x_0^*, \theta_0^*)$, then the optimal incentives are $\arg\min\{\varphi(x_0^*, \theta_0^*, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$. But since we do not know the initial conditions, the above result shows that both (3.12) and $\arg\min\{\varphi(\hat{x}_{0,T}, \hat{\theta}_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$ are reasonable approximations. In fact, we can show a stronger result for the solution generated by 2SSA.

**Theorem 3.8.** *Note that* $\arg\min\{\varphi(x_0^*, \theta_0^*, \{\pi_i\}_{i=0}^{T+n}) | \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$ *is the set of optimal solutions under the agent's true initial conditions* $(x_0^*, \theta_0^*)$. *If Assumptions 3.1–3.8 hold, then we have that*

$$\operatorname{dist}\Big(\pi_{2SSA}(T), \arg\min\{\varphi(x_0^*, \theta_0^*, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}\Big) \xrightarrow{p} 0 \tag{3.16}$$

*as* $T \to \infty$*, for any* $\pi_{2SSA}(T)$ *returned by 2SSA. Note* $\operatorname{dist}(x, B) = \inf_{y \in B} \|x - y\|$.

The full proof of this theorem can be found in Appendix B.1, but here we will provide some intuition for the proof. This result follows by applying Proposition 3.7 and Theorem 4.3 from (Vogel and Lachout, 2003a).

This result suggests that any solution returned by 2SSA is asymptotically included within the set of optimal incentives computed for the agent's true initial conditions. Restated, the result says 2SSA provides a set of incentives that are asymptotically optimal. This is a non-obvious result because in general pointwise-convergence of a sequence of stochastic optimization problems is not sufficient to ensure convergence of the minimizers of the sequence of optimization problems to the minimizer of the limiting optimization problem. Rockafellar and Wets (2009) provide an example in their Figure 7–1 that demonstrates this possible lack of convergence of minimizers.

### 3.4.2 Policy Calculation With Multiple Agents

We next study the general setting where the coordinator designs incentives for a large group of agents. We let $\mathcal{A}$ be the set of agents, and the quantities corresponding to a specific agent $a \in \mathcal{A}$ are denoted using subscript $a$. Now suppose that at time $T$ the coordinator measures $\{\tilde{x}^a_{t_i}\}_{i=0}^{n^a_x}, \{\tilde{y}^a_{\tau_i}\}_{i=0}^{n^a_u}$ for all agents $a \in \mathcal{A}$. One approach to designing incentives is by solving:

$$\min\Big\{\mathbb{E}[\Phi(x^a_0, \theta^a_0, \{\pi^a_i\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) | \{\tilde{x}^a_{t_i}\}_{i=0}^{T^a_x}, \{\tilde{y}^a_{\tau_i}\}_{i=0}^{T^a_u}, \{\pi^a_i\}_{i=0}^{T} \text{ for } a \in \mathcal{A}] \ \Big| \ \{\{\pi^a_i\}_{i=T+1}^{T+n} \text{ for } a \in \mathcal{A}\} \in \Omega\Big\} \quad (3.17)$$

Here, $\Phi : \mathcal{X}^{\#\mathcal{A}} \times \theta^{\#\mathcal{A}} \times \Omega \to \mathbb{R}$ is a joint loss function that depends on the behavior of all agents. For the settings we are interested in, this loss function has a separable structure.

**Assumption 3.9.** Loss $\Phi$ is additively $\Phi(x^a_0, \theta^a_0, \{\pi^a_i\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) = \sum_{a\in\mathcal{A}} \varphi^a(x^a_0, \theta^a_0, \{\pi^a_i\}_{i=0}^{T+n})$ or multiplicatively separable $\Phi(x^a_0, \theta^a_0, \{\pi^a_i\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) = \prod_{a\in\mathcal{A}} \varphi^a(x^a_0, \theta^a_0, \{\pi^a_i\}_{i=0}^{T+n})$.

Without loss of generality, we assume $\Phi$ is additively separable since we can obtain similar results for the case of multiplicative separability by taking the logarithm of $\Phi$. We also make an assumption that states $\Omega$ is decomposable in a simple way.

**Assumption 3.10.** There exist a finite set $V = \{v_1, v_2, \ldots\}$ with vector-valued, sets $S_v \subseteq \Pi^n$ for $v \in V$, and a vector-valued constant $\beta$ such that

$$\Omega = \Big\{\{\pi^a_i\}_{i=T+1}^{T+n} \text{ for } a \in \mathcal{A} : y^a_v \in \{0,1\}, \textstyle\sum_{v\in V} y^a_v = 1 \text{ for } a \in \mathcal{A}, \sum_{a\in\mathcal{A}}\sum_{v\in V} v{\cdot}y^a_v \leq \beta \\ \{\pi^a_i\}_{i=T+1}^{T+n} \in S_v \text{ if } y^a_v = 1\Big\}. \quad (3.18)$$

Moreover, the sets $S_v$ are compact sets that are representable by a finite number of mixed integer linear constraints.

The underlying idea of this assumption is that the set $V$ describes a vector of discrete elements that can be used as incentives, and $\beta$ is the vector-valued budget on

**Algorithm 2** Adaptive Behavioral Multi-Agent Algorithm (ABMA)

```
Require: $\{\tilde{x}^a_{t_i}\}_{i=0}^{n_x^a}, \{\tilde{u}^a_{\tau_i}\}_{i=0}^{n_u^a}, \{\pi^a_i\}_{i=0}^{T}$ for $a \in \mathcal{A}$
for all $a \in \mathcal{A}$ do
    compute $(\hat{x}^a_{0,T}, \hat{\theta}^a_{0,T}) = \text{argmax}_{(x_0,\theta_0)}\, \psi_T(x_0, \theta_0)$
    for all $v \in V$ do
        set $\pi^a_v \in \arg\min\{\varphi^a(\hat{x}_{0,T}, \hat{\theta}_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in S_v\}$
        set $\phi^a_v = \varphi^a(\hat{x}^a_{0,T}, \hat{\theta}^a_{0,T}, \pi^a_v)$
    end for
end for
compute $y := \{y^a_v :\ a, v \in \mathcal{A} \times \mathcal{V}\}$:
    $$\begin{aligned} y \in \text{argmin}\ & \textstyle\sum_{a\in\mathcal{A}} \sum_{v \in V} \phi^a_v \cdot y^a_v \\ \text{s.t.}\ & \textstyle\sum_{a\in\mathcal{A}} \sum_{v \in V} v \cdot y^a_v \leq \beta \\ & \textstyle\sum_{v \in V} y^a_v = 1 \text{ for } a \in \mathcal{A} \\ & y^a_v \in \{0, 1\} \quad \text{for } a, v \in \mathcal{A} \times \mathcal{V} \end{aligned}$$
for all $a \in \mathcal{A}$ and $v \in V$ do
    set $\pi^a_{ABMA}(T) = \pi^a_v$ if $y^a_v = 1$
end for
return $\pi^a_{ABMA}(T)$ for $a \in \mathcal{A}$
```

the discrete incentives. When the discrete incentives are fixed at $v$, the set $S_v$ keeps the discrete incentives fixed and describes the feasible set of continuous incentives.

Even with these assumptions on separability and decomposibility, solving (3.17) is difficult because it is a BMIP with $\#\mathcal{A}$ MILP's in the lower level. Thus, we develop an adaptive algorithm (based on the 2SSA algorithm) for optimizing incentives for multiple agents. We first summarize our algorithm, and demonstrate that it only requires solving a small number of computable MILP's. Next we prove this algorithm provides a set of incentives that are asymptotically optimal as time advances.

### Adaptive Algorithm for Multiple Agents

We design incentives for multiple agents with the Adaptive Behavioral Multi-Agent Algorithm (ABMA) presented in Algorithm 2. The main idea behind this method is to use the assumptions on $\Phi$ and $\Omega$ to decompose the initial problem into $\#\mathcal{A}$ sub-problems that solve a single agent problem, and a single master problem that combines these solutions into a global optimum across all agents. Because of the assumptions on $\Omega$, each sub-problem can be further decomposed into $\#V$ sub-problems. For each sub-problem, we use the 2SSA algorithm to solve the $\#\mathcal{A} \cdot \#V$ sub-problems; however, we do not explicitly call the 2SSA algorithm because it is more efficient to solve the MAP estimator once and then solve the incentive design problem for each single agent. Our first result concerns the computability of this algorithm.

**Proposition 3.9.** If Assumptions 3.1–3.10 hold, then the main computational steps of the ABMA algorithm involve solving a total of $\#\mathcal{A} \cdot (\#V + 1)$ MILP's and 1 ILP.

The full proof of this proposition can be found in Appendix B.1 but here we will provide some intuition for this proof. The result can be directly calculated by noting that Step 2 of ABMA can be computed by solving a single MILP. Similarly, steps 4 and 5 and also be computed by solving a single MILP each. Then Step 8 requires computing a pure ILP by construction. Since the remaining steps of ABMA do not require solving optimization problems, we only need to count the amount of times each step is repeated thus yielding the result.

This means the ABMA algorithm performs incentive design for the multi-agent case by solving $\#\mathcal{A} \cdot (\#V + 1) + 1$ MILP's, which is significantly less challenging than solving a BMIP with $\#\mathcal{A}$ MILP's in the lower level as would be required to solve (3.17).

The ABMA algorithm also has an alternative interpretation, and to better understand this consider the following feasibility problem:

$$\begin{aligned}
&\Phi(\overline{x}_0^a, \overline{\theta}_0^a, \{\overline{\pi}_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) = \\
&\min_{\{x_t^a, u_t^a, \theta_t^a, \pi_t^a\}_{t=0}^{T+n}, \forall a \in \mathcal{A}} \quad \Phi(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) \\
&\qquad \text{s.t.} \quad \begin{array}{l} u_t^a \in \operatorname{argmax} f(x_{t+1}^a, u, \theta_t^a, \pi_t^a) \\ \qquad \text{s.t. } x_{t+1}^a = h(x_t^a, u), u \in \mathcal{U} \end{array} \qquad \forall a, 0 \leq t \leq T + n - 1, \\
&\qquad\qquad \theta_{t+1}^a = g(x_t^a, u_t^a, \theta_t^a, \pi_t^a) \qquad \forall a, 0 \leq t \leq T + n - 1, \\
&\qquad\qquad x_t^a \in \mathcal{X}^a, \theta_t^a \in \Theta, \pi_t^a \in \Pi \qquad \forall a, 0 \leq t \leq T + n - 1, \\
&\qquad\qquad x_0^a = \overline{x}_0^a, \theta_0^a = \overline{\theta}_0^a, \{\pi_t^a\}_{t=0}^{T+n} = \{\overline{\pi}_t^a\}_{t=0}^{T+n} \; \forall a, \\
&\qquad\qquad \{\{\pi_t^a\}_{t=T+1}^{T+n} \text{ for } a \in \mathcal{A}\} \in \Omega.
\end{aligned} \tag{3.19}$$

Our first result concerns regularity properties of the above written feasibility problem.

**Proposition 3.10.** If Assumptions 3.1–3.10 hold, then $\Phi(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A})$ is lower semicontinuous in its arguments, and $\min\{\Phi(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n}$ for $a \in \mathcal{A}\} \in \Omega\}$ can be formulated as a MILP for all fixed values of $(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T}) \in \mathcal{X} \times \Theta \times \Pi^{T+1}$ for $a \in \mathcal{A}$.

The full proof of this proposition can be found in Appendix B.1 but here we will provide some intuition for the proof. We can obtain the first result by first showing that problem (3.19) can be reformulated as a MILP. Since $(\overline{x}_0^a, \overline{\theta}_0^a, \{\overline{\pi}_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A})$ appear in affine terms in this new MILP formulation, this means that the value function of the optimization problem is lower semicontinuous thus proving the first result. The second result follows by showing that the desired optimization problem is equivalent to (3.19) but with removal of the constraints $\pi_t^a = \overline{\pi}_t^a$ for $t = T+1, \ldots, T+n$.

The optimization problem (3.19) and the above result provide an alternative interpretation of the ABMA algorithm, which is formalized by the next corollary.

**Corollary 3.11.** If Assumptions 3.1–3.10 hold, then the solution of $\min\{\Phi(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n}$ for $a \in \mathcal{A}\} \in \Omega\}$, is given by the ABMA algorithm but with Step 2 replaced with the step: set $(\hat{x}_{0,T}^a, \hat{\theta}_{0,T}^a) = (x_0^a, \theta_0^a)$.

This is straightforward from the reformulation shown in (B.12).

Thus, though (3.19) is a large MILP, the assumptions we have made allow us to decompose the solution of this problem into a series of substantially smaller MILP's.

### Asymptotic Optimality of ABMA

The optimization problem in (3.19) is a useful construction because it can also be used to compute the optimal set of incentives. If each agent's true initial conditions $(x_0^{*,a}, \theta_0^{*,a})$ were known, then an optimal solution belongs to $\arg\min\{\Phi(x_0^{*,a}, \theta_0^{*,a}, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n}$ for $a \in \mathcal{A}\} \in \Omega\}$. More importantly, we have the following relationship to the solutions of the ABMA algorithm:

**Theorem 3.12.** *Note that* $\arg\min\{\Phi(x_0^{*,a}, \theta_0^{*,a}, \{\pi_i^{*,a}\}_{i=0}^{T+n}$ *for* $a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n}$ *for* $a \in \mathcal{A}\} \in \Omega\}$ *is the set of optimal solutions under the agents' true initial conditions* $(x_0^{*,a}, \theta_0^{*,a})$. *If Assumptions 3.1–3.8 hold, then we have that*

$$\begin{aligned}&\mathrm{dist}\Big(\{\pi_{ABMA}^a(T) \text{ for } a \in \mathcal{A}\},\\ &\arg\min\{\Phi(x_0^{*,a}, \theta_0^{*,a}, \{\pi_i^a\}_{i=0}^{T+n} \textit{ for } a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n} \textit{ for } a \in \mathcal{A}\} \in \Omega\}\Big) \xrightarrow{p} 0 \quad (3.20)\end{aligned}$$

*as* $T \to \infty$, *for any* $\pi_{ABMA}^a(T)$ *returned by ABMA. Recall that* $\mathrm{dist}(x, B) = \inf_{y \in B}\|x - y\|$.

A complete proof for this theorem can be found in Appendix B.1 but here we will provide some intuition for the proof. Since the MAP estimator is consistent and by Corollary 3.6 we have that $\Phi$ is lower semicontinuous, this means that $\Phi(\hat{x}_0^a, \hat{\theta}_0^a, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A})$ is a lower semicontinuous approximation to $\Phi(x_0^{*,a}, \theta_0^{*,a}, \{\pi_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A})$. Hence applying Corollary 3.11 and Theorem 4.3 from (Vogel and Lachout, 2003a) we obtain the desired result.

Thus any solution returned by ABMA is asymptotically included within the set of optimal incentives computed for the agents' true initial conditions. Restated, the above result says ABMA provides a set of incentives that are asymptotically optimal. This is a non-obvious result because in general pointwise-convergence of a sequence of stochastic optimization problems is not sufficient to ensure convergence of the minimizers of the sequence of optimization problems to the minimizer of the limiting optimization problem. Rockafellar and Wets (2009) provide an example that demonstrates this possible lack of convergence of minimizers.

## 3.5 Computational Experiments: Weight Loss Program Design

We have completed computational experiments applying the tools and techniques developed in this chapter that form a specific implementation of a behavioral analytics framework. We compare several approaches, including ours, for designing incentives for multiple myopic agents to the problem described in Section 3.1.1 of designing personalized behavioral incentives for a clinically-supervised weight loss program. The first step of our behavioral analytics approach is to construct a behavioral model of individuals in weight loss programs. We describe the data source used for the simulations, and then summarize our behavioral model (Aswani et al., 2016) for individuals participating in such loss programs. To demonstrate the second and third steps of our framework, we simulate a setting in which behavioral incentives chosen using our ABMA algorithm are evaluated against behavioral incentives computed by (intuitively-designed) adaptive heuristics. Both our implementation of a behavioral analytics framework and the heuristic provide adaptation by recomputing the incentives at regular intervals as more data is collected from each individual. Our metric for comparison is the number of individuals who achieve clinically significant weight loss (i.e., a 5% reduction in body weight) at the end of the program. We also compare the percentage of weight loss for individuals who do not achieve clinically significant weight loss in order to better understand how clinical visits are allocated by the different methods. We conclude by performing a sensitivity analysis of design choices for the second and third steps of our behavioral analytics framework.

### 3.5.1 mDPP Program Trial Data Source

Our computational experiments used data from the mDPP trial (Fukuoka et al., 2015). This was a randomized control trial (RCT) that was conducted to evaluate the efficacy of a 5 month mobile phone based weight loss program among overweight and obese adults at risk for developing type 2 diabetes. This program was adapted from the

Diabetes Prevention Program (DPP) (Diabetes Prevention Program Research Group, 2002, 2009), but the number of in person clinical visits was reduced from 16 to 6 per person, and group exercise sessions were replaced with a home based exercise program to reduce costs. Sixty one overweight adults were randomized into an active control group which only received an accelerometer (n=31) or a treatment group which receive the mDPP mobile app plus the accelerometer and clinical office visits (n=30). Changes in primary and secondary outcomes for the trial were promising. The treatment group lost an average of 6.2 ± 5.9 kg (-6.8% ± 5.7%) between baseline and the 5 month follow up while the control group gained 0.3 ± 3.0 kg (0.3% ± 5.7 %) ($p < 0.001$). The treatment group's steps per day increased by 2551 ± 4712 compared to the control group's decrease of 734 ± 3308 steps per day ($p < 0.001$). Additional details on demographics and other treatment parameters are available in (Fukuoka et al., 2015). The data available from the mDPP trial includes step data (from accelerometer measurements), body weight data (which was measured at least twice a week every week and recorded in the mobile app by individuals in the treatment group, as well as measured three times in a clinical setting at baseline, 3 month, and 5 month), and demographic data (i.e., age, gender, and height of each individual). We note that this data matches the assumptions in Section 3.2.

### 3.5.2 Summary of Behavioral Model

We construct a behavioral model for each individual participating in the weight loss program. Using the terminology and notation of Section 3.2, the system state of each individual $x_t$ is their body weight on day $t$ which we denote as $w_t$, and their decisions $u_t = (f_t, s_t)$ on day $t$ are how many calories they consume $f_t$ and how many steps they walk $s_t$. The behavioral incentives $\pi_t = (g_t, d_t)$ provided to an individual on day $t$ consist of (numeric) step goals $g_t$ and an indicator $d_t$ equal to one if a clinical visit was scheduled for that day. The motivational state (or type) is $\theta_t := (s_b, f_{b,t}, F_{b,t}, p_t, \mu, \delta, \beta)$. The state $s_b$ captures the individual's baseline preference for number of steps taken each day, while $f_{b,t}, F_{b,t}$ capture the individual's short term and long term caloric intake preference, respectively. The variable $p_t$ captures the disutility an individual experiences from not meeting a step goal. The last set of motivational states describe the individual's response to behavioral incentives. The states $\beta, \delta$ describe the amount of change in the individual's caloric consumption and physical activity preferences, respectively, after undergoing a single clinical visit. The state $\mu$ describes the self efficacy effect (Bandura, 1998, Conner and Norman, 1996) from meeting exercise goals.

The utility function of an individual on day $t$ is given by $f(w_{t+1}, f_t, s_t; \theta_t, \pi_t) = -w_{t+1}^2 - r_s(s_t - s_b)^2 - r_f(f - f_{b,t})^2 + p_t(s_t - g_t)^-$. In our past modeling work (Aswani et al., 2016), we found that the predictions of this behavioral model were relatively insensitive to the value of $r_f, r_s$. And our numerical experiments in (Aswani et al.,

2016) found that fixing the value of $r_f, r_s$ to be the same for each individual provided accurate predictions. And so we assume that $r_f, r_s$ is a fixed and known constant in our numerical experiments here. The temporal dynamics for an individual's system and motivational states are the dynamics of an individual's type by:

$$\begin{aligned} w_{t+1} &= a \cdot w_t + b \cdot s_t + c \cdot f_t + k && (3.21)\\ F_{b,t+1} &= (1-\alpha) \cdot F_{b,t} + \alpha \cdot f_{b,t} && (3.22)\\ f_{b,t+1} &= \gamma \cdot (f_{b,t} - F_{b,t}) + F_{b,t} - \beta \cdot d_n && (3.23)\\ p_{t+1} &= \gamma \cdot p_t + \delta \cdot d_t + \mu \cdot \mathbb{1}(s_t \geq g_t). && (3.24) \end{aligned}$$

Equation (3.21) is a "calories in minus calories out" description of weight change, and a standard physiological formula (Mifflin et al., 1990) is used to compute the values of $a, b, c, k$ based on the demographics of the individual. Equation (3.22) models the long term caloric intake preference as an exponential moving average of the short term caloric intake preferences. We found that the predictions for different individuals were relatively insensitive to the value of $\alpha$, and so in our numerical experiments we assume $\alpha$ is known and fixed to a value satisfying $\alpha < 1$. In (3.23), we model the dynamics of baseline food consumption as always tending towards their initial value unless perturbed by a clinical visit. In (3.24) we model the tendency for meeting the step goal as tending towards zero unless there is a clinical visit or the individual has met the previous exercise goal, which increases their self efficacy and makes the individual more likely to meet their step goal in the future. In both (3.23) and (3.24), $\gamma < 0$ is assumed to be a known decay factor since we found that predictions were relatively insensitive to the value of $\gamma$ (Aswani et al., 2016). Note that these temporal dynamics and utility functions satisfy the assumptions in Section 3.2.

For the MLE and MAP calculations, we assumed that step and weight data were measured with zero-mean noise distributed according to a Laplace distribution with known variance. We found that predictions of the behavioral model estimated when assuming the noise had a Gaussian distribution were of the same quality as those estimated when assuming Laplace noise, and so we assume Laplace noise so that the MAP and MLE problems can be formulated as MILP's (as shown in Section 3.2). Furthermore, the prior distribution used for the MAP calculation for each individual was a histogram of the MLE estimates of all the other individual's parameters. Note that this form for a prior distribution can be expressed using integer linear constraints (Aswani et al., 2016). The complete MILP formulations for MAP and MLE are provided in the appendix.

### 3.5.3 Weight Loss Program Design

Since the majority of implementation costs for weight loss programs are due to clinical visits, the clinician's design problem is to maximize the expected number of individuals who reduce their weight by a clinically significant amount (i.e., 5% reduction in body weight). The clinician is able to personalize the step goals for each individual, and can change the number and timing of clinical visits for each individual. However, there is a budget constraint on the total number of visits that can be scheduled across all individuals. This constraint captures the costliness of clinical visits.

We optimize the weight loss program using our ABMA algorithm to implement the second and third steps of our behavioral analytics framework. This requires choosing a loss function for each individual, and Figure 3.1 shows three choices that we considered. These three losses make varying tradeoffs between achieving the primary health outcome of number of individuals with clinically significant weight loss (i.e., 5% weight loss) at the end of the program versus the secondary health outcome of maximizing weight loss of individuals who were not able to achieve 5% weight loss. The first choice of a loss function is the *step loss*, which is given by

$$\varphi = \begin{cases} -1, & \text{if } w_{T+n} \leq 0.95 \cdot \tilde{w}_0 \\ 0, & \text{otherwise} \end{cases} \tag{3.25}$$

This discontinuous choice of a loss function gives minimal loss to 5% or more reduction in body weight and maximal loss to less than 5% reduction in body weight. The second choice of a loss function is the *hinge loss*, which is given by

$$\varphi = \begin{cases} -1, & \text{if } w_{T+n} < 0.95 \cdot \tilde{w}_0 \\ -0.2 \cdot (w_{T+n}/\tilde{w}_0 - 1), & \text{if } 0.95 \cdot \tilde{w}_0 \leq w_{T+n} \leq \tilde{w}_0 \\ 0, & \text{if } w_{T+n} > \tilde{w}_0 \end{cases} \tag{3.26}$$

This continuous choice of a loss function gives minimal loss to 5% or more reduction in body weight, maximal loss to less than 0% reduction in body weight, and an intermediate loss for intermediate reductions in body weight. The third choice of a loss function is the *time-varying hinge loss*, which is given by

$$\varphi = \begin{cases} -1, & \text{if } \frac{w_{T+n}}{\tilde{w}_0} < 0.95 - \frac{0.05}{\log T} \\ 10(\log T)(\frac{w_{T+n}}{\tilde{w}_0} - (0.95 + \frac{0.05}{\log T})), & \text{if } 0.95 - \frac{0.05}{\log T} \leq \frac{w_{T+n}}{\tilde{w}_0} \leq 0.95 + \frac{0.05}{\log T} \\ 0, & \text{if } \frac{w_{T+n}}{\tilde{w}_0} > 0.95 + \frac{0.05}{\log T} \end{cases} \tag{3.27}$$

Much like the hinge loss (3.26), it promotes intermediate amounts of weight loss that might not meet the 5% threshold of clinically significant weight loss. However, as

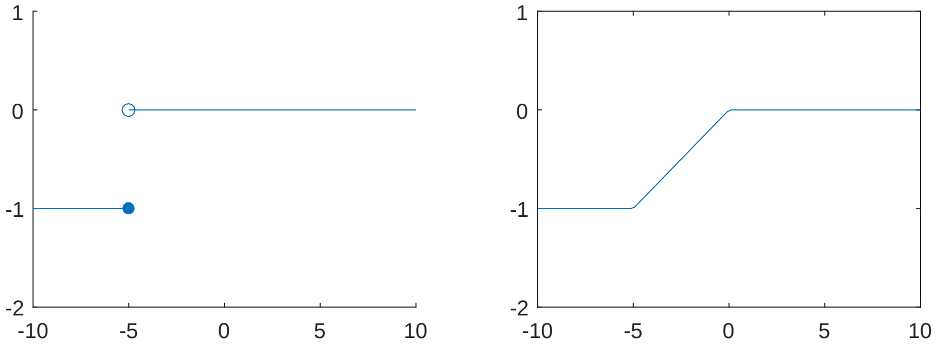


Figure 3.1: The left plot shows the step loss function (3.25), and the right plot shows the hinge loss function (3.26). The $x$-axis on both plots is $100 \cdot (x_{T+n}/\tilde{x}_0 - 1)$, which is percent reduction in body weight.

more data is collected it approaches the step loss (3.25) to reflect a higher degree of confidence in the estimated parameters. Thus, this choice of loss function can be considered an intermediate between the hinge (3.26) and step (3.25) losses. There is one computational note. Since these losses are non-decreasing, we can modify Step 4 of the ABMA algorithm to instead minimize the body weight of each individual and then compose the body weight with the loss function.

For the purpose of comparing various program designs through simulations, we considered three additional designs for the weight loss program. We used an adaptive heuristic to design the weight loss program: Clinical visits were scheduled towards the end of the treatment at least one week apart, with more visits given to individuals who were closer to meeting the weight loss goal of a 5% weight reduction based on their latest observed weight, and step goals were set to be a 10% increase over a linear moving average of the individual's observed step count over the prior week. The second design was a "do nothing" plan where individuals were given exactly one clinical visit after two weeks, and their step goals were a constant 10,000 steps each day. The third design was the original design of the mDPP trial: Clinical visits were scheduled on predetermined days during the treatment after 2, 4, 6, 19, 14, 18, and 20 weeks of the trials. The first two weeks of this design did not contain any clinical visits or exercise goals but instead served as an initialization period. After the first two weeks, exercise goals increased 20% each week, starting with a 20% increase over the average number of steps taken by individuals during the the two week initialization period. The exercise goals were capped at a maximum of 12,000 steps a day. Since the adaptive heuristic and ABMA are both adaptive, we recalculated both at the beginning of each month of the treatment and allowed both adaptive methods a 2 week initialization time similar to the mDPP trial.

### 3.5.4 Simulation Comparison

We compared the six different program designs using simulations of a weight loss program with a five month duration and with 30 individuals participating. Each individual in the simulation followed our behavioral model, and the parameters corresponding to the behavioral model for each individual were chosen to be those estimated by computing the MLE using the data from the mDPP trial. Since we also wanted to test how these different designs account for missing data, we assumed that the data available to each algorithm would be limited to days of the mDPP study where a particular individual reported their weight and steps. Since the adaptive heuristic and our behavioral analytics framework are both adaptive, we recalculated the program design at the beginning of each month of the program (by re-runnning the heuristic calculations and rerunning the ABMA algorithm) and allowed both adaptive methods a two-week initialization time similar to the design of the program in the mDPP trial. All simulations were run using MATLAB on a laptop computer with a 2.4GHz processor and 16GB RAM. The Gurobi solver (Gurobi Optimization, 2015) in conjunction with the CVX toolbox for MATLAB (Grant and Boyd, 2014) were used to perform the initial estimation of the individual parameters, compute designs for the weight loss program, and perform simulations of each design.

Figure 3.2 compares the primary outcome of interest to clinicians, which is the number of individuals that achieve clinically significant weight loss (i.e., 5% or more reduction in body weight). We repeated the simulations for our behavioral analytics framework and the adaptive heuristic under different constraints on the total number of clinical visits that could be allocated to individuals. The $x$-axis of Figure 3.2 is the average number of clinical visits provided to individuals. The horizontal line at 18 is the number of individuals who achieved 5% weight loss in the actual mDPP trial, in which each individual received 7 clinical visits. Figure 3.2 shows that all forms of behavioral analytics program and adaptive heuristic program designs outperform the "do nothing" policy. Furthermore, our behavioral analytics approach and the adaptive heuristic achieve results comparable to the original mDPP program design but with significantly less resources (i.e., less clinical visits). The simulations predict that using our behavioral analytics approach in which ABMA has a step (3.25) or time-varying hinge loss (3.27) to design the weight loss program can provide health outcomes comparable to current clinical practice while using only 40-60% of the resources (i.e., clinical visits) of current practice. In contrast, the adaptive heuristic would require 80-95% of resources (i.e., clinical visits) to attain health outcomes comparable to current clinical practice. This suggests that appropriate choice of the loss function for our ABMA algorithm, as part of behavioral analytics approach, to personalize the design of a weight loss program could increase capacity (in terms of the number of individuals participating in the program for a fixed cost) by up to 60%, while achieving comparable

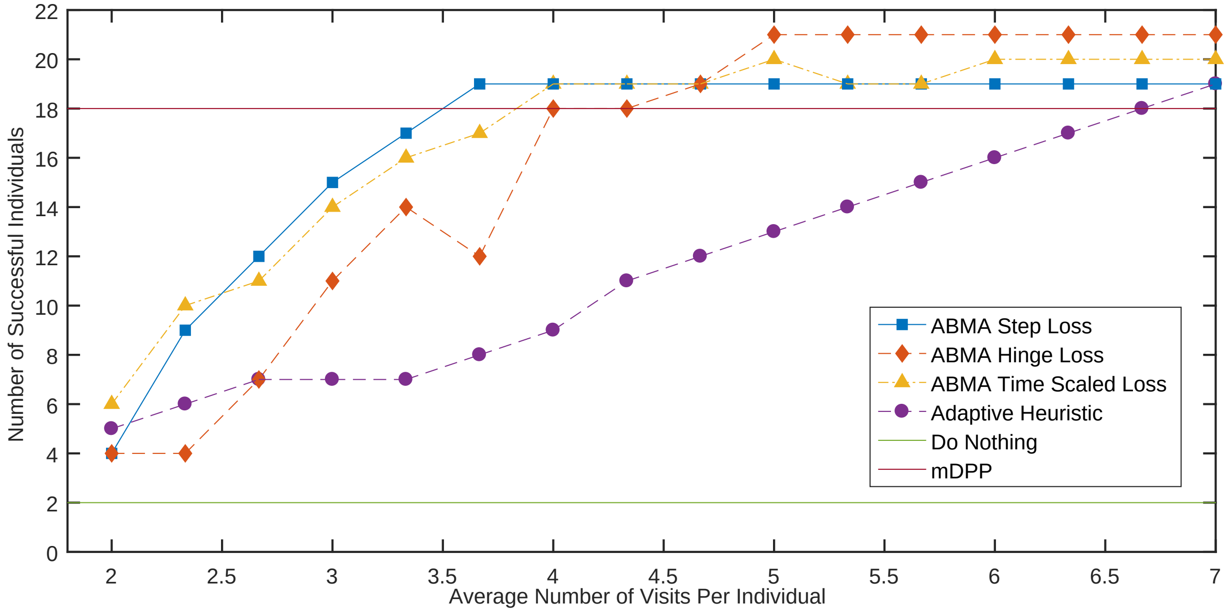


Figure 3.2: Comparison of different program design methods with respect to number of successful individuals (i.e., lost 5% or more body weight)

health outcomes.

Figure 3.3 compares the different program designs using a secondary outcome of interest to clinicians of the average amount of weight loss of individuals who did not successfully achieve 5% weight loss. The original treatment plan of mDPP and the "do nothing" treatment plan slightly outperform the adaptive program designs at certain clinical visit budgets. This effect however is mainly due to these static plans not identifying individuals who are on the cusp of achieving 5% weight loss but might still achieve around 3-4% weight loss, while both adaptive program designs allocate clinical visits to these individuals and ensure they reach the weight loss goal of 5% weight loss. Restated, the lower weight loss of unsuccessful individuals under the behavioral analytics treatment plans is an artifact of the improved success rate of the behavioral analytics plans in helping individuals achieve 5% weight loss. This effect is further exemplified in the region of between an average of 2.8-4.2 visits per individual, where we see that individuals who were not successful in achieving 5% weight loss in the behavioral analytics treatment plans on average lose more weight then those under the heuristic policy, while in the region of an average of 5.6-7 visits per individual we see that individuals under the heuristic treatment lose more weight. The effect in this last region is mainly due to the behavioral model used in our behavioral analytics framework, which is more effective at identifying the individuals who would most benefit from additional clinical visits. Therefore, more resources are spent on

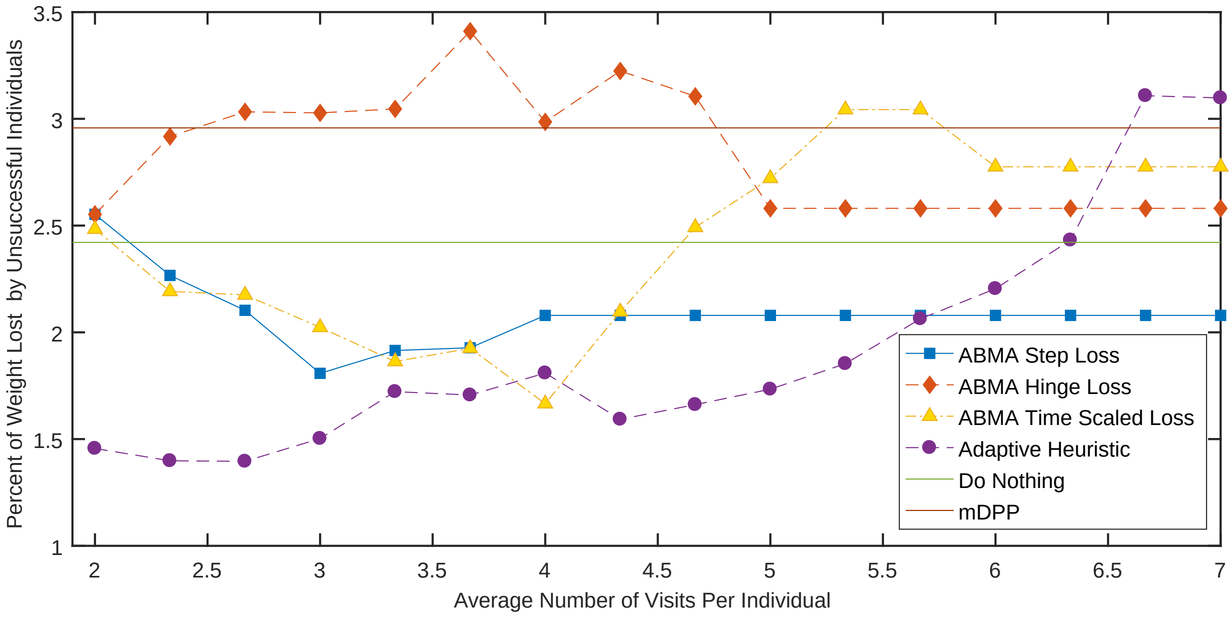


Figure 3.3: Comparison of different program design methods with respect to the percent weight lost by unsuccessful individuals (i.e., lost less than 5% body weight)

individuals who could potentially reach their 5% weight loss goal, while the adaptive heuristic uses these resources in a less effective manner.

Figures 3.2 and 3.3 demonstrate a tradeoff between the primary and secondary outcomes, and the various loss functions provide different tradeoffs. Note the line for the step loss (3.25) is the first to achieve a primary outcome comparable to that of the original mDPP trial while fluctuating relatively little in terms of the secondary outcome. This matches intuition that the step loss function (3.25) is focused on ensuring individuals achieve 5% weight loss while not being concerned with their final weight. On the other hand, the line for the hinge loss (3.26) lags behind the other behavioral analytics policies in achieving comparable primary outcomes to the mDPP trial while having an extremely effective secondary outcome. These results follow our intuition that this loss favors intermediate weight loss over achieving clinically significant weight loss. Finally, the line corresponding to the time-varying hinge loss function (3.27) has a clear transition at an average of 5 visits per individual from favoring the primary outcome to the secondary outcome. This behavior indicates that using such time scaling leads to interventions that focus on primary outcomes when resources are constrained but also accounts for secondary outcomes when resources are less scarce. Such behavior may be useful for implementing a behavioral analytics approach when the relative abundance of resources is not known *a priori*.

### 3.5.5 Computational Performance and Sensitivity Analysis

The simulation experiments assumed that treatment plans were updated at the beginning of each month by re-running the second and third steps of our behavioral analytics approach (through applying the ABMA algorithm), and so we conducted a sensitivity analysis to examine the effect of updating the program design more or less frequently. Figures 3.4 and 3.5 compare the health outcomes of using a program designed by our behavioral analytics framework with a time-varying hinge loss (3.27), where the treatment was recalculated once every two weeks, once a month, and once every two months. These results show that recomputing the treatment plan with lesser or higher frequency does not significantly impact the efficacy of the resulting treatment. This indicates that for practical implementation, the statistical convergence rate of estimated parameters in our behavioral model is sufficiently fast that it would suffice to rerun the weight loss program design algorithm at most once a month.

We also conducted time-benchmarks for the sub-problems involved in computing 2SSA, MAP, and ABMA, which are the algorithms comprising the second and third steps of our implementation o a behavioral analytics framework. The results of the time-benchmarks are summarized in Tables B.1,B.2, and B.3 in the appendix. On average, solving all sub-problems took 17s per individual. This is promising for practical implementation, particularly because each sub-problem calculation can be performed in parallel for each size constraint of the clinical visit schedule (from 1 to 7 visits). The results show that computation time increases with respect to the number of available visits and data available in the treatment plan calculation. However, the calculation times still remain below 30s on average per individual for each step of the program calculation. This would imply that our methodology for weight loss program calculation is suitable for large scale program design since the program design would be updated at most once every month.

## 3.6 Conclusion

In this chapter, we develop a *behavioral analytics* framework for multi-agent systems in which a single coordinator provides behavioral or financial incentives to a large number of myopic agents. Our framework is applicable in a variety of settings of interest to the operations research community, including the design of demand-response programs for electricity consumes, the personalized design of a weight loss program, and adaptive logistics allocation for franchises. The framework we develop involves the definition of a behavioral model, the estimation of model parameters, and the optimization of incentives. We show (among other results) that under mild assumptions, the incentives computed by our approach converge to the optimal incentives that would be computed

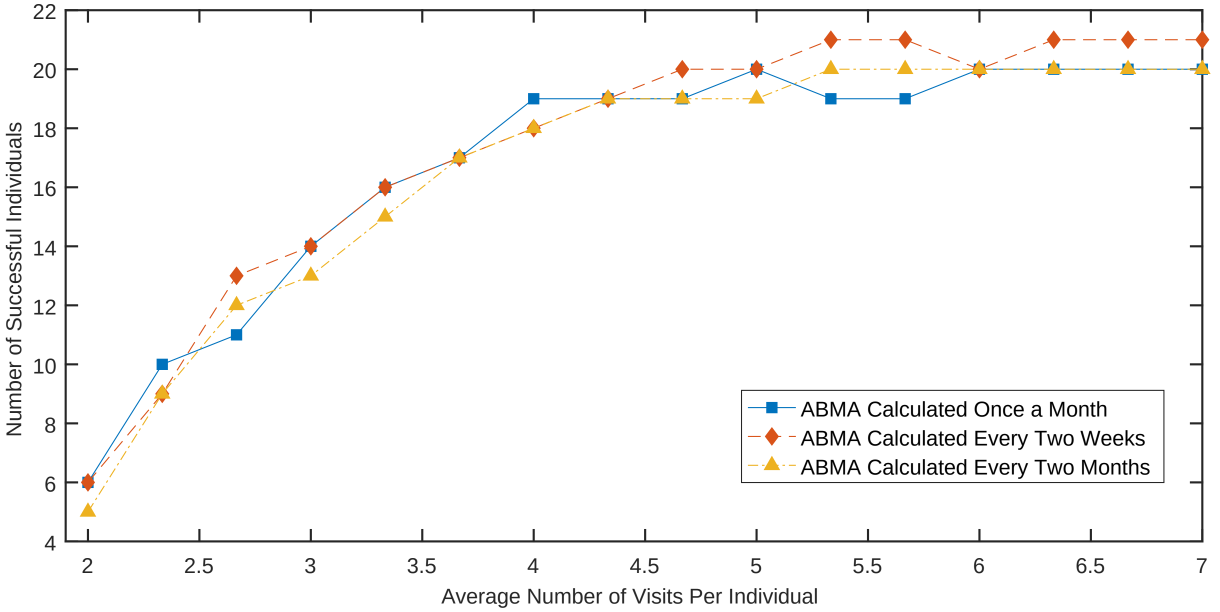


Figure 3.4: Comparison of calculation schedules and their effects on the number of successful individuals (i.e., lost 5% or more body weight)

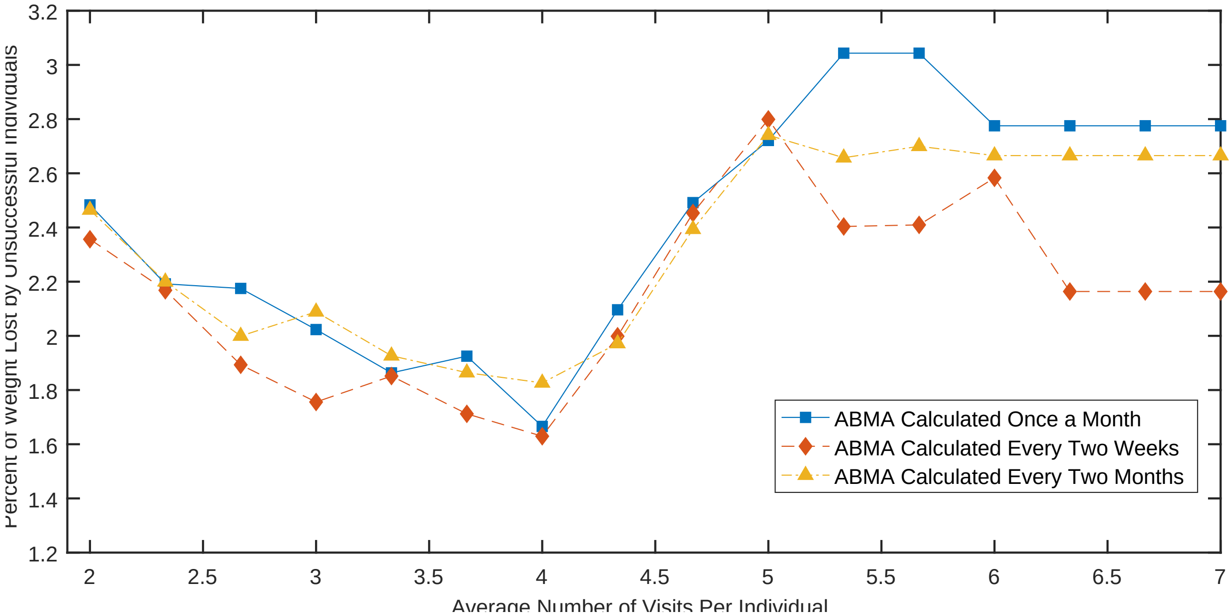


Figure 3.5: Comparison of calculation schedules and their effects on the weight lost by unsuccessful individuals (i.e., lost less than 5% body weight)

knowing full information about the agents. We evaluated our approach for personalizing the design of a weight loss program, and showed via simulation that our approach can improve outcomes with reduced treatment cost.

# Chapter 4

# Interventions with Cheep and Frequent Decisions

## 4.1 Introduction

Multi-armed bandits are commonly used to model sequential decision-making in settings where there is a set of actions that can be chosen at each time step, each action provides a stochastic reward, and the distribution for the reward provided by each action is initially unknown. The problem of constructing a policy for sequentially choosing actions in multi-armed bandits requires balancing *exploration* versus *exploitation*, the tradeoff between selecting what is believed to be the action that provides the best reward and choosing other actions to better learn about their underlying distributions. Bandit models have been applied in a variety of healthcare settings (Thompson, 1933, Wang and Gupta, 2011, Bastani and Bayati, 2015b, Schell et al., 2016). For instance, Bastani and Bayati (2015b) considered the problem of selecting drugs to give to a patient from a set (where each drug is an action) in order to treat a specific disease (the reward is the improvement in patient health in response to the drug); the bandit policy asymptotically identifies the optimal drug for that particular patient. Other common applications involve online advertising (Agrawal and Goyal, 2013, Johari et al., 2015), where selecting an ad to show is an action and the reward is the total number (from a large population) of viewers who click on the ad, as well as in various supply chain settings (Afèche and Ata, 2013, Ban and Rudin, 2014, Caro and Gallien, 2007).

However, most bandit models assume that the distribution for the reward provided by each action is constant over time. This is a reasonable assumption in a large number of applications, such as the ones described above. However, many applications involve actions that are applied to a single individual, where the rewards depend upon behavioral responses of the individual to the applied actions. In these behavioral settings, the response to a particular action is not generally stationary. Frequent selection of a particular action will lead to habituation to that action by the individual, and the reward for that action will decrease each time it is selected. For example, repeatedly showing the same ad to a single individual may cause the ad to become less effective in soliciting a response from that individual. Furthermore, another complimentary phenomenon can also occur; refraining for a period of time from showing a particular ad to a single individual may cause the ad to become more effective when reintroduced.

Most techniques for designing policies for decision-making for multi-armed bandits apply to the setting where the rewards for each action are stationary. However, designing a policy without considering the non-stationarity of a system (when the system is in fact non-stationary) often leads to poor results in terms of maximizing rewards (Besbes et al., 2014, Hartland et al., 2006) because policies eventually converge to a stationary policy. The problem of designing policies for bandit models with non-stationarity has been studied in specific settings, but approaches in the literature are either computationally intractable, or the settings analyzed are not flexible enough to capture the habituation and recovery phenomenon described above. The aim of this chapter is to propose a flexible bandit model that is able to effectively model habituation and recovery, and to present an approach for designing an effective policy for this bandit model.

### 4.1.1 Literature Review

Data-driven decision-making can be categorized into batch formulations and online formulations. Batch formulations (Aswani et al., 2016, Mintz et al., 2017a, Ban and Rudin, 2014, Ban, 2015, Bertsimas et al., 2014) use a large amount of data to estimate a predictive model and then use this model for optimization. Adaptation to new data occurs by reestimating the predictive model, which is done periodically after a specified amount of additional data is collected.

On the other hand, online formulations involve constructing a policy that is updated every time a new data point is collected. Bandit models are a particularly important example of online formulations, and there has been much work on constructing policies for stationary bandits. Approaches for designing policies for stationary bandits include those using upper confidence bounds (Auer et al., 2002a, Chang et al., 2005, Bastani and Bayati, 2015b), Thompson sampling (Thompson, 1933, Russo and Roy, 2014, 2016, Agrawal and Goyal, 2013), Bayesian optimization (Frazier and Wang, 2016, Xie and Frazier, 2013, Xie et al., 2016), knowledge gradients (Ryzhov and Powell, 2011, Ryzhov et al., 2012), robust optimization (Kim and Lim, 2015), and adversarial optimization (Auer et al., 2002b, Agrawal et al., 2014, Koolen et al., 2014, 2015).

Restless bandits are a notable class of bandit models that capture non-stationarity, because choosing any single action causes the rewards of potentially all the actions to change. Though dynamic programming (Liu and Zhao, 2010, Whittle, 1988), approximation algorithms (Guha et al., 2010), and mathematical programming (Bertsimas and Nino-Mora, 1994, Bertsimas and Niño-Mora, 2000, Caro and Gallien, 2007) have been proposed as tools for constructing policies in this setting, the problem of computing an optimal policy for restless bandits is PSPACE-complete (Papadimitriou and Tsitsiklis, 1999), meaning that designing policies that are approximately optimal is difficult.

Another related research stream designs policies for non-stationary multi-armed bandits with specific structures. For instance, model-free approaches have been proposed (Besbes et al., 2014, 2015, Garivier and Moulines, 2008, Anantharam et al., 1987) for settings with bounded variations, so that rewards of each action are assumed to change abruptly but infrequently. These policies have been shown to achieve $\mathcal{O}(\sqrt{T \log T})$ suboptimality. Recently, there has been interest in studying more structured non-stationary bandits. Two relevant examples are Adjusted Upper Confidence Bounds (A-UCB) and rotting bandits (Bouneffouf and Féraud, 2016, Levine et al., 2017), where each action has a set of unknown but stationary parameters and a set of known non-stationary parameters that characterize its reward distribution. Policies designed for these settings achieve $\mathcal{O}(\log T)$ suboptimality, but these settings are unable to capture the habituation and recovery phenomenon that is of interest to us.

### 4.1.2 ROGUE Bandits

In this chapter, we define the ROGUE (reducing or gaining unknown efficacy) bandit model, which can capture habituation and recovery phenomenon, and then we design a nearly-optimal policy for this model. ROGUE bandits are appropriate for application domains where habituation and recovery are important factors for system design; we present two such examples, in online advertising and personalized healthcare, below.

#### Personalized Healthcare-Adherence Improving Interventions

One hundred fifty minutes of moderate-intensity aerobic physical activity each week has been shown to reduce the risk of cardiovascular disease, other metabolic disorders, and certain types of cancers (Committee et al., 2008, Friedenreich et al., 2010, Sattelmair et al., 2011, Lewis et al., 2017). However, maintaining this level of moderate intensity activity is challenging for most adults. As such, proper motivation through providing daily exercise goals and encouragement has been found to be effective in helping patients succeed in being active (Fukuoka et al., 2011, 2014, 2015).

In recent years, there has been an increased rate of adoption of fitness applications and wearable activity trackers, making it easier and less costly to implement physical activity programs (PwC, 2014). These trackers and mobile applications record daily activity, communicate activity goals, and send motivational messages. Despite these digital devices having collected a large amount of personal physical activity data, many of the most popular activity trackers provide static and non-personalized activity goals and messages to their users (Rosenbaum, 2016). Furthermore, the choice of motivational messages sent to users may have significant impact on physical activity, because if users receive similar messages too frequently they may become habituated and not respond with increased activity, while seldom sent messages may better increase activity due to their novelty and diversity. Because the ROGUE bandits can model habituation and recovery of rewards for different actions, we believe they present a useful framework for the design of policies that choose which messages to send to users based on data consisting of what messages they received each day and the corresponding amounts of physical activity on those days.

Personalized healthcare has been extensively studied in the operations literature. Aswani et al. (2016), Mintz et al. (2017a) explore the use of behavioral analytics to personalize diet and exercise goals for clinically supervised weight loss interventions in an offline setting. Markov decision processes have also been used for decision-making in personalized healthcare (Ayer et al., 2015, Mason et al., 2013, Deo et al., 2013, Kucukyazici et al., 2011, Leff et al., 1986, Wang and Gupta, 2011, Gupta and Wang, 2008, Savelsbergh and Smilowitz, 2016, Schell et al., 2016). In contrast to bandit models where only the reward for the prescribed action can be observed, these methods broadly assume that the full state of the system can be observed, and thus do not require statistical estimation. Additionally, various multi-armed bandit approaches (Bastani and Bayati, 2015b, Wang and Gupta, 2011) have also been proposed for healthcare problems where habituation and recovery are not significant factors.

### Online Content Creation and Advertising

Online advertising is one of the fastest-growing industries in the US. In fact, as of 2016, US Internet advertising spending has increased to over \$72.5 billion, surpassing the amount spent on TV ads (Richter, 2017). However, as this form of advertising becomes more prevalent, advertisers have been struggling to ensure that ads retain there effectiveness.This has been attributed to Internet users being habituated by impersonal and standardized ads (Goldfarb and Tucker, 2014, Portnoy and Marchionini, 2010) which are rarely varied. For these reasons, there has been significant interest in the operations literature in creating automated systems that can utilize user-level data to better target and customize ads (Ghose and Yang, 2009, Goldfarb and Tucker, 2011). In particular, since the effect of a no-longer-effective advertisement may recover after a user has not seen it for some period of time, incorporating recovery and habituation dynamics into advertising models could yield more effective advertising campaigns.

In general, multi-armed bandit models have been proposed to model online advertising, where each action corresponds to a different type of advertisement, and the reward is equivalent to either a conversion or a click from a prospective consumer. Several approaches have been used to design such an ad targeting system, including adversarial and stochastic multi-armed bandit models (Bertsimas and Mersereau, 2007, Chen et al., 2013, Kleinberg et al., 2008, Liu et al., 2010, Yi-jun et al., 2010), and online statistical testing (Johari et al., 2015). However, while some of these approaches use contextual data to better serve ads to individuals, they are still designed under assumptions of stationarity. As a result, these approaches will lead to policies that show duplicated ads to individuals, which can potentially causing habituation, whereas other ads that might have recovered efficacy may not be served at all. In contrast, ROGUE Bandit models can explicitly consider the time-varying efficacy each type of ad, and thus directly capture user habituation to a specific ad, and track the recovery of efficacy of a particular ad for a specific individual.

### 4.1.3 Outline

In Section 4.2, we formally introduce the ROGUE bandit model. To the best of our knowledge, this is the first work where a non-stationary bandit model has been defined that is able to capture habituation and recovery phenomenon, and is at the same time amenable to the design of nearly-optimal policies. Because the ROGUE bandit is a general model, we describe two specific instantiations: the ROGUE generalized linear model and the ROGUE agent.

Next, in Section 4.3 we analyze the problem of estimating the parameters of a single action. We present a statistical analysis of maximum likelihood estimation (MLE) for a single action, and use empirical process theory to derive finite sample bounds for the convergence of parameters estimates. Specifically, we show that the MLE estimates converge to the true parameters at a $1/\sqrt{T}$ rate.

Section 4.4 describes an upper-confidence bound policy for ROGUE bandits, and we call this policy the ROGUE-UCB algorithm. The main result of this section is a rigorous $\mathcal{O}(\log T)$ bound on the suboptimality of the policy in terms of regret, the difference between the reward achieved by the policy and the reward achieved by an optimal policy. Our $\mathcal{O}(\log T)$ bound is significant because this is the optimal rate achievable for approximate policies in the stationary case (Lai and Robbins, 1985). We prove our bound using methods from the theory of concentration of measure.

We conclude with Section 4.5, where we introduce a "tuned" version of ROGUE-UCB and then conduct numerical experiments to compare the efficacy of our ROGUE-UCB algorithm to other policies that have been developed for bandit models. Our experiments involve two instantiations of ROGUE bandit models. First, we compare different bandit policies using a ROGUE generalized linear bandit to generate data. Second, we compare different bandit policies using a ROGUE agent to generate data, where the parameters of this bandit model are generated using data from a physical activity and weight loss clinical trial (Fukuoka et al., 2014). This second experiment specifically addresses the question of how to choose an optimal sequence of messages to send to a particular user in order to optimally encourage the user to increase physical activity, and it can be interpreted as a healthcare-adherence improving intervention. Our experiments show that ROGUE-UCB outperforms all other considered bandit policies, and that it achieves logarithmic regret, in contrast to other bandit algorithms that achieve linear regret.

## 4.2 Defining Reducing or Gaining Unknown Efficacy (ROGUE) Bandits

This section first describes the stationary multi-armed bandit (MAB) model, in order to emphasize modeling differences in comparison to our ROGUE bandit model that is introduced in this section. Our goal in defining ROGUE bandits is to have a model that can capture specific non-stationary phenomena found in behavioral applications, and so we next formally introduce the model elements of ROGUE bandits. To provide better intuition about ROGUE bandits, we also present two specific instantiations of a ROGUE bandit that incorporate different behavioral effects.

### 4.2.1 Stationary MAB Model

The stationary MAB is a setting where there is a finite set of actions $\mathcal{A}$ that can be chosen at each time step $t$, each action $a \in \mathcal{A}$ provides a stochastic reward $r_a$ with distribution $\mathbb{P}_{\theta_a}$, and the parameters $\theta_a \in \Theta$ for $a \in \mathcal{A}$ are constants that are initially unknown but lie in a known compact set $\Theta$. The problem is to construct a policy for sequentially choosing actions in order to maximize the expected reward. More specifically, let $\pi_t \in \mathcal{A}$ be the action chosen at time $t = 1, \dots, T$. Then the policy consists of functions $\pi_t(r_{\pi_1}, \dots, r_{\pi_{t-1}}, \pi_1, \dots, \pi_{t-1}) \in \mathcal{A}$ that depend on past rewards and actions. For notational convenience, we will use $\Pi = \{\pi_t(\cdot)\}_{t=1}^T$ to refer to the policy. In this notation, the problem of constructing an optimal policy to maximize expected reward can be written as $\max_{\Pi \in \mathcal{A}^T} \sum_{t=1}^T \mathbb{E} r_{\pi_t}$. Note that certain regularity is needed from the distributions to ensure this maximization problem is well-posed. One common set of assumptions is that the distributions $\mathbb{P}_{\theta_a}$ for $a \in \mathcal{A}$ are sub-Gaussian, and that the reward distributions are all independent.

For the stationary MAB, we can define an optimal action $a^* \in \mathcal{A}$, which is any action such that $\mathbb{E} r_{a^*} \geq \mathbb{E} r_a$ for all $a \in \mathcal{A}$. The benefit of this definition is it allows us to reframe the policy design problem in terms of minimizing the cumulative expected regret $\mathbb{E} R_\Pi(T) = \mathbb{E}[T r_{a^*} - \sum_{i=1}^T r_{\pi_t}]$, where the quantity $r_{a^*} - r_{\pi_t}$ is known as the regret at time $t$. Observe that minimizing $\mathbb{E} R_\Pi(T)$ is equivalent to maximizing $\sum_{t=1}^T \mathbb{E} r_{\pi_t}$. It has been shown by Gittins (1979) that an index policy is optimal for the stationary MAB. Since these indexing policies are difficult to compute, other approximate policies have been proposed (Lai and Robbins, 1985, Auer et al., 2002a). Some of the most common policies use upper confidence bounds (Auer et al., 2002a, Garivier and Cappé, 2011), which take actions optimistically based on estimates of the parameters $\theta_a$. Unfortunately, it has been shown that these index policies and upper confidence bound policies can have arbitrarily bad performance in a non-stationary setting (Hartland et al., 2006, Besbes et al., 2014).

### 4.2.2 Reducing or Gaining Unknown Efficacy (ROGUE) Bandits

A disadvantage of the stationary MAB is that it does not allow rewards to change over time in response to previous actions, and this prevents the stationary MAB model from being able to capture habituation or recovery phenomena. Here, we define ROGUE bandits that can describe such behavior. The ROGUE bandit is a setting where there is a finite set of actions $\mathcal{A}$ that can be chosen at each time step $t$, each action $a \in \mathcal{A}$ at time $t$ provides a stochastic reward $r_{a,t}$ that has a sub-Gaussian distribution $\mathbb{P}_{\theta_a, x_{a,t}}$ with expectation $\mathbb{E}r_{a,t} = g(\theta_a, x_{a,t})$ for a bounded function $g$, the parameters $\theta_a \in \Theta$ for $a \in \mathcal{A}$ are constants that are initially unknown but lie in a known compact, convex set $\Theta$, and each action $a \in \mathcal{A}$ has a state $x_{a,t}$ with nonlinear dynamics

$$x_{a,t+1} = \mathrm{proj}_{\mathcal{X}}(A_a x_{a,t} + B_a \pi_{a,t} + K_a) = h(x_{a,t}, \pi_{a,t}), \tag{4.1}$$

where $\pi_{a,t} = \mathbf{1}[\pi_t = a]$, $\mathcal{X}$ is a known compact, convex set, $A_a, B_a, K_a$ are known matrices and vectors, and $x_{a,0}$ is initially unknown for $a \in \mathcal{A}$. Note that the effect of the projection in the dynamics (4.1) is to act as a saturator of the state, so that the state does not become unbounded.

The problem is to construct a policy for sequentially choosing actions in order to maximize the expected reward. Observe that the ROGUE bandit model is non-stationary since the reward distributions depend upon previous actions. This makes the problem of designing policies more difficult than that of designing policies for the stationary MAB. More specifically, let $\pi_t \in \mathcal{A}$ be the action chosen at time $t = 1, \ldots, T$. Then the policy consists of functions $\pi_t(r_{\pi_1}, \ldots, r_{\pi_{t-1}}, \pi_1, \ldots, \pi_{t-1}) \in \mathcal{A}$ that depend on past rewards and actions. For notational convenience, we will use $\Pi = \{\pi_t(\cdot)\}_{t=1}^T$ to refer to the policy. In this notation, the problem of constructing an optimal policy to maximize expected reward can be written as

$$\max\nolimits_{\Pi \in \mathcal{A}^T} \{\textstyle\sum_{t=1}^T g(\theta_{\pi_t}, x_{\pi_t,t}) : x_{a,t+1} = h_a(x_{a,t}, \pi_{a,t}) \text{ for } a \in \mathcal{A},\ t \in \{0, ..., T-1\}\}. \tag{4.2}$$

This can be reframed as minimizing expected cumulative regret (Besbes et al., 2014, Garivier and Moulines, 2008, Bouneffouf and Féraud, 2016): Unlike the stationary

MAB, we cannot define an optimal action, but rather must define an optimal policy $\Pi^* = \{\pi_t^*(\cdot)\}_{t=0}^T$, which can be thought of as an oracle that chooses the optimal action at each time step. Then the problem of designing an optimal policy is equivalent to minimizing $R_\Pi(T) = \sum_{t=1}^T r_{\pi_t^*,t} - r_{\pi_t,t}$ subject to the state dynamics defined above.

### 4.2.3 Technical Assumptions on ROGUE Bandits

In this chapter, we will design a policy for ROGUE bandits that follow the assumptions described below:

**Assumption 4.1.** The rewards $r_{a,t}$ are conditionally independent given $x_{a,0}, \theta_a$ (or equivalently the complete sequence of $x_{a,t}, \pi_t$ and $\theta_a$).

This assumption states that for any two time points $t, t'$ such that $t \neq t'$ we have that $r_{a,t}|\{x_{a,t}, \theta\}$ is independent of $r_{a,t'}|\{x_{a,t'}, \theta\}$, and it is a mild assumption because it is the closest analogue to the assumption of independence of rewards in the stationary MAB.

**Assumption 4.2.** The reward distribution $\mathbb{P}_{\theta,x}$ has a log-concave probability density function (p.d.f.) $p(r|\theta, x)$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$.

This assumption provides regularity for the reward distributions, and is met by many common distributions (e.g., Gaussian and Bernoulli).

Now define $f(\cdot)$ to be $L$-Lipschitz continuous if $|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_2$ for all $x_1, x_2$ in the domain of $f$. Our next assumption is on the stability of the above distributions with respect to various parameters.

**Assumption 4.3.** The log-likelihood ratio $\ell(r; \theta', x', \theta, x) = \log \frac{p(r|\theta',x')}{p(r|\theta,x)}$ associated with the distribution family $\mathbb{P}_{\theta,x}$ is $L_f$-Lipschitz continuous with respect to $x, \theta$, and $g$ is $L_g$-Lipschitz continuous with respect to $x, \theta$.

This assumption ensures that if two sets of parameters are close to each other in value then the resulting distributions will also be similar. We make the following additional assumption about the functional structure of the reward distribution family:

**Assumption 4.4.** The reward distribution $\mathbb{P}_{\theta,x}$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$ is sub-Gaussian with parameter $\sigma$, and either $p(r|\theta, x)$ has a finite support or $\ell(r; \theta', x', \theta, x)$ is $L_p$-Lipschitz with respect to $r$.

This assumption (or a similar type of regularity) is needed to ensure that sample averages are close to their means, and it is satisfied by many distributions (e.g., a Gaussian location family with known variance).

Last, we impose conditions on the dynamics for the state of each action:

**Assumption 4.5.** We assume $\|A_a\|_2 \leq 1$ for all $a \in \mathcal{A}$, where $\|\cdot\|_2$ is the usual matrix 2-norm.

This assumption is needed to ensure the states of each action do not change too quickly, and it is equivalent to assuming that the linear portion of the dynamics is stable.

### 4.2.4 Instantiations of ROGUE Bandits

The above assumptions are general and apply to many instantiations of ROGUE bandit models. To demonstrate the generality of these assumptions, we present two particular instances of ROGUE bandit models.

#### ROGUE Agent

Our first instantiation of a ROGUE bandit model consists of a dynamic version of a principal-agent model (Stackelberg, 1952, Radner, 1985, Laffont and Martimort, 2002, Mintz et al., 2017a), which is a model where a principal designs incentives to offer to an agent who is maximizing an (initally unknown to the principal) utility function that depends on the incentives. In particular, consider a setting with a single (myopic) agent to whom we would like to assign a sequence of behavioral incentives $\pi_t \in \mathcal{A}$, and the states $x_{a,t}$ and parameters $\theta_a$ are scalars. Given a particular incentive $\pi_t$ at time $t$, the agent responds by maximizing the (random) utility function

$$r_t = \operatorname{argmax}_{r\in[0,1]} -\tfrac{1}{2}r^2 - (c_{a,t} + \textstyle\sum_{a\in\mathcal{A}} x_{a,t}\pi_{t,a})r, \tag{4.3}$$

where for fixed $a \in \mathcal{A}$ we have that $c_{a,t}$ are i.i.d. random variables with a distribution $\mathbb{P}_{\theta_a}$ such that $\mathrm{Var}(c_{a,t}) = \sigma^2(\theta_a) < \infty$ and $\sigma^2 : \mathbb{R} \to \mathbb{R}_+$ is invertible. Moreover, the state dynamics are

$$x_{a,t+1} = \mathrm{proj}_{\mathcal{X}}(\alpha_a x_{a,t} + b_a(1 - \pi_{a,t}) - k_a), \tag{4.4}$$

which is of form (4.1) with $B_a = -k_a$, $K_a = b_a - k_a$, and $A_a = \alpha_a$. Note the distribution of $r_t$ is fully determined by $x_{a,t}, \theta_a, \{\pi_k\}_{k=0}^t$, which means the rewards satisfy Assumption 4.1.

We can further analyze the above ROGUE agent model. Solving the agent's optimization problem (4.3) gives

$$r_t|\{x_{a,t}, \theta_a\} = \begin{cases} 0 & \text{if } \ c_{a,t} \leq -x_{a,t}, \\ 1 & \text{if } \ c_{a,t} \geq 1 - x_{a,t}, \\ c_{a,t} + x_{a,t} & \text{otherwise} \end{cases} \tag{4.5}$$

We can express the distribution of $r_t|\{x_{a,t}, \theta_a\}$ in terms of the cumulative distribution function (c.d.f.) $F(\cdot)$ and p.d.f. $f(\cdot)$ of $c_{a,t}$:

$$p(r_t|\{x_{a,t}, \theta_a\}) = F(-x_{a,t})\delta(r_t) + (1 - F(1 - x_{a,t}))\delta(1 - r_t) \\ + f(r_t - x_{a,t})\mathbf{1}[r_t \in (0, 1)]. \quad (4.6)$$

Though $p(r_t|\{x_{a,t}, \theta_a\})$ is not an absolutely continuous function, it satisfies Assumptions 4.2 and 4.3, whenever $c_t$ has a log-concave p.d.f. that is Lipschitz continuous, if we interpret the above probability measure $p(r_t|\{x_{a,t}, \theta_a\})$ as a p.d.f.

### ROGUE Generalized Linear Model (GLM)

Dynamic logistic models and other dynamic generalized linear models (McCullagh, 1984, Filippi et al., 2010) can be interpreted as non-stationary generalizations of the classical (Bernoulli reward) stationary MAB (Gittins, 1979, Lai and Robbins, 1985, Garivier and Cappé, 2011). Here, we further generalize these models: Consider a setting where $r_{a,t}|\{\theta_a, x_{a,t}\}$ is an exponential family with mean parameter

$$\mu_{a,t} = \mathbb{E}r_t = g(\alpha_a^T\theta_a + \beta_a^T x_{a,t}), \quad (4.7)$$

for known vectors $\alpha_a, \beta_a$, where the action states $x_{a,t}$ have the dynamics (4.1). In this situation, we can interpret $g(\cdot)$ as a link function of a generalized linear model (GLM). For example, if $g$ is a logit function, then this model implies the rewards have a Bernoulli distribution with parameter

$$\mu_{a,t} = \frac{1}{1 + \exp(-(\alpha_a^T\theta_a + \beta_a^T x_{a,t}))}. \quad (4.8)$$

For the logistic case, the $r_{a,t}$ is bounded and satisfies Assumptions 4.1-4.2. These assumptions are also satisfied if $r_{a,t}$ can be linked to a truncated exponential family distribution restricted to $[0, 1]$, meaning if the p.d.f. of $r_{a,t}|\{x_{a,t}, \theta_a\}$ is

$$\frac{h(r)}{F(1) - F(0)} \exp\left(T(r)g(\alpha_a^T\theta_a + \beta_a^T x_{a,t}) - A(\alpha_a^T\theta_a + \beta_a^T x_{a,t})\right), \quad (4.9)$$

where $T(r)$ is a sufficient statistic. If instead we consider sub-Gaussian exponential families with infinite support, Assumption 4.4 is satisfied if the sufficient statistic of the GLM is Lipschitz or bounded with respect to $r$. While we will mainly consider one-dimensional rewards (i.e., $r_{a,t} \in \mathbb{R}$), we note that this framework can also be extended to vector and array dynamic GLM's.

## 4.3 Parameter Estimation for ROGUE Bandits

Our approach to designing a policy for ROGUE bandits will involve generalizing the upper confidence bound policies (Auer et al., 2002a, Chang et al., 2005, Bastani and Bayati, 2015b) that have been developed for variants of stationary MAB's. As per the name of these policies, the key step involves constructing a confidence bound for the parameters $\theta_a, x_{a,0}$ characterizing the distribution of each action $a \in \mathcal{A}$. This construction is simpler in the stationary case because the i.i.d. structure of the rewards allows use of standard Chernoff-Hoeffding bounds (Wainwright, 2015), but we can no longer rely upon such i.i.d. structure for ROGUE bandits which are fundamentally non-stationary. This is because in ROGUE bandits the reward distributions depend upon states $x_{a,t}$, and so the structure of ROGUE bandits necessitates new theoretical results on concentration of measure in order to construct upper confidence bounds for the relevant parameters.

For this analysis, let the variables $\{r_{a,t}\}_{t=1}^{T}$ be the observed rewards for action $a \in \mathcal{A}$. It is important to note that the $r_{a,t}$ here are no longer random variables, but are rather the actual observed values. Since the reward distributions for each action are mutually independent by the dynamics (4.1), we can study the estimation problem for only a single action. Specifically, consider the likelihood $p(\{r_{a,t}\}_{t\in\mathcal{T}_a}|\theta_a, x_{a,0})$, where $\mathcal{T}_a \subset \{1, ..., T\}$ is the set of times when action $a$ was chosen (i.e., $\pi_t = a$ for $t \in \mathcal{T}_a$). Let $n(\mathcal{T}_a)$ denote the cardinality of the set $\mathcal{T}_a$. Using Assumption 4.1, the likelihood can be expressed as

$$p(\{r_{a,t}\}_{t\in\mathcal{T}_a}|\theta_a, x_{a,0}) = \prod_{t\in\mathcal{T}_a} p(r_{a,t}|\theta_a, x_{a,t}) \prod_{t\in\mathcal{T}_a} p(x_t|\theta_a, x_{a,t_-}). \tag{4.10}$$

where $t_- = \max\{s \in \mathcal{T}_a : s < t\}$ is the latest observation before time $t$. Note the MLE of $\theta_a, x_{a,0}$ is $(\hat{\theta}_a, \hat{x}_{a,0}) \in \text{argmax} \prod_{t\in\mathcal{T}_a} p(r_{a,t}|\theta_a, x_{a,t}) \prod_{t\in\mathcal{T}_a} p(x_t|\theta_a, x_{a,t_-})$. Observe that by (4.1), the one step likelihood $p(x_t|\theta_a, x_{a,t-1})$ is a degenerate distribution with all probability mass at $x_{a,t}$, by perpetuation of the dynamics (4.1) with initial conditions $x_{a,t-1}$. Thus we can express the MLE as the solution to the constrained optimization problem

$$\begin{aligned}(\hat{\theta}_a, \hat{x}_{a,0}) = \arg\min\{-\textstyle\sum_{t\in\mathcal{T}_a} \log p(r_{a,t}|\theta_a, x_{a,t}) :& \\ x_{a,t+1} = h(x_{a,t}, \pi_{a,t}) \text{ for } t \in \{0, \dots, T\}\},&\end{aligned} \tag{4.11}$$

where we have also taken the negative logarithm of the likelihood (4.10). In this section, we will consider concentration properties of the solution to the above optimization problem. If $\theta^*_a, x^*_{0,a}$ for $a \in \mathcal{A}$ are the true parameter values of a ROGUE Bandit model, then we show that

**Theorem 4.1.** *For any constant $\xi > 0$ we have*

$$P\left(\frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \hat{\theta}_a, \hat{x}_{a,0}) \leq \xi + \frac{c_f(d_x, d_\theta)}{\sqrt{n(\mathcal{T}_a)}}\right) \geq 1 - \exp\left(\frac{-\xi^2 n(\mathcal{T}_a)}{2L_p^2\sigma^2}\right) \tag{4.12}$$

*where*

$$c_f(d_x, d_\theta) = 8L_f \operatorname{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{d_x+d_\theta}} L_f \operatorname{diam}(\mathcal{X} \times \Theta)\sqrt{\pi(d_x + d_\theta)} \tag{4.13}$$

*is a constant that depends upon $d_x$ (the dimensionality of $\mathcal{X}$) and $d_\theta$ (the dimensionality of $\Theta$), and $D_{a,\pi_1^T}(\theta_a, x_{a,0}||\theta_a', x_{a,0}') = \sum_{t\in\mathcal{T}_a} D_{KL}(\mathbb{P}_{\theta_a,x_{a,t}}||\mathbb{P}_{\theta_a',x_{a,t}'})$ is the trajectory KullbackLeibler (KL) divergence between two different initial conditions.*

### 4.3.1 Conceptual Reformulation of MLE

Our analysis begins with a reformulation of the MLE that removes the constraints corresponding to the dynamics (4.1) through repeated composition of the function $h_a$ defining the dynamics (4.1).

**Proposition 4.2.** Let $\theta_a^* \in \Theta$ and $x_{a,0}^* \in \mathcal{X}$ for $a \in \mathcal{A}$ be the true underlying parameters of the system, then the MLE is given by

$$(\hat{\theta}_a, \hat{x}_{a,0}) = \underset{\theta_a, x_{a,0} \in \Theta\times\mathcal{X}}{\operatorname{argmin}} \frac{1}{n(\mathcal{T}_a)} \sum_{t\in\mathcal{T}_a} \log\frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \tag{4.14}$$

where the notation $h_a^k$ represents the repeated functional composition of $h_a$ with itself $k$ times, and $\pi_1^t$ is the sequence of input decisions from time 1 to time $t$.

The complete proof for this proposition is found in Appendix C.1, and here we provide a sketch of the proof. Observe that this formulation is obtained by first adding constant terms equal to the likelihood of the true parameter values to the objective function and dividing by the total number of observations (which does not change the optimal solution), and then composing our system dynamics and writing them as explicit functions of the initial conditions. In practice, this reformulation is not practical to solve since clearly $\theta_a^*, x_{a,0}^*$ are not known *a priori* and the composite function $h_a^t$ may have a complex form. However, for theoretical analysis this reformulation is quite useful, since for fixed $\theta_a, x_{a,0}$ taking the expected value of the objective under $\mathbb{P}_{\theta_a^*,x_{a,0}^*}$ yields

$$\begin{aligned}\mathbb{E}_{\theta_a^*,x_{a,0}^*}\frac{1}{n(\mathcal{T}_a)} \sum_{t\in\mathcal{T}_a} \log\frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} &= \frac{1}{n(\mathcal{T}_a)} \sum_{t\in\mathcal{T}_a} D_{KL}(\mathbb{P}_{\theta_a^*,x_{a,t}^*}||\mathbb{P}_{\theta_a,x_{a,t}}) \\ &= \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^*||\theta_a, x_{a,0}). \end{aligned} \tag{4.15}$$

Essentially, we have reformulated the MLE problem in terms of minimizing the KL divergence between the trajectory distribution of potential sets of parameters to the trajectory distribution of the true parameter set. Since we have clear interpretation for the expectation of our objective function we can now proceed to compute concentration inequalities.

### 4.3.2 Uniform Law of Large Numbers for ROGUE Bandits

Since our estimates are computed by solving an optimization problem, a pointwise law of large numbers is insufficient for our purposes since such a result would not be strong enough to imply convergence of the optimal solutions. To obtain proper concentration inequalities we must consider a uniform law of large numbers for the MLE problem.

**Theorem 4.3.** *For any constant $\xi > 0$ we have*

$$P\left(\sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \left| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \right.\right.$$
$$\left.\left. - \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \theta_a, x_{a,0}) \right| > \xi + \frac{c_f(d_x, d_\theta)}{\sqrt{n(\mathcal{T}_a)}} \right) \leq \exp\left( \frac{-\xi^2 n(\mathcal{T}_a)}{2L_p^2 \sigma^2} \right) \quad (4.16)$$

*where*

$$c_f(d_x, d_\theta) = 8L_f \operatorname{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{d_x + d_\theta}} L_f \operatorname{diam}(\mathcal{X} \times \Theta)\sqrt{\pi(d_x + d_\theta)} \quad (4.17)$$

*is a constant.*

We will prove this result in several steps, the first of which uses the following lemma:

**Lemma 4.1.** *Consider the mapping*

$$\varphi\Big(\{r_t\}_{t=1}^{n(\mathcal{T}_a)}\Big) = \sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \left| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \right.$$
$$\left. - \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \theta_a, x_{a,0}) \right|. \quad (4.18)$$

*The mapping $\varphi$ is $L_p$-Lipschitz with respect to $\{r_t\}_{t=1}^{n(\mathcal{T}_a)}$.*

A detailed proof is provided in Appendix C.1, and the main argument of the proof relies on the preservation of Lipschitz continuity through functional composition and pointwise maximization. This result is necessary since showing that objective value variations are bounded is a prerequisite for the formalization of concentration bounds. Next we consider the Lipschitz constant of the log-likelihood with respect to the parameters.

**Lemma 4.2.** *For any $r \in \mathbb{R}$, $\bar{\theta} \in \Theta$, $\bar{x} \in \mathcal{X}$, define the function $\ell : \Theta \times \mathcal{X} \times \{1, ..., T\} \to \mathbb{R}$ such that $\ell(\theta, x, t) = \log \frac{p(r|\bar{\theta}, h_a^t(\bar{x}, \bar{\theta}, \pi_1^t))}{p(r|\theta, h_a^t(x, \theta, \pi_1^t))}$. Then for fixed $t$, the function $\ell$ is Lipshitz with constant $L_f$. Moreover, for all $(x, \theta) \in \mathcal{X} \times \Theta$ and for all $t, t' \in \{1, ..., T\}$ we have that $|\ell(\theta, x, t) - \ell(\theta, x, t')| \leq L_f \operatorname{diam}(\mathcal{X})$, where $\operatorname{diam}(\mathcal{X}) = \max_{x \in \mathcal{X}} \|x\|_2$.*

The result of this lemma can be derived using a similar argument to that of Lemma 4.1, by noting that the dynamics are bounded and Lipschitz, and then applying Assumption 4.3. The full proof of this lemma is in Appendix C.1. Next we show the expected behavior of $\pi$ is bounded.

**Lemma 4.3.** *Let $\varphi$ be defined as in Lemma 4.1. Then $\mathbb{E}\varphi(\{r_t\}_{t=1}^{n(\mathcal{T}_a)}) \leq \frac{c_f(d_x, d_\theta)}{\sqrt{n(\mathcal{T}_a)}}$, where*

$$c_f(d_x, d_\theta) = 8L_f \operatorname{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{d_x + d_\theta}} L_f \operatorname{diam}(\mathcal{X} \times \Theta)\sqrt{\pi(d_x + d_\theta)}. \quad (4.19)$$

The result of this lemma is derived by first using a symmetrization argument to bound the expectation by a Rademacher average and then using metric entropy bounds to derive the final result, and a complete proof is found in Appendix C.1. Additional insight into these results is provided by the following remarks:

**Remark 4.1.** The result of Lemma 4.3 implies that $\mathbb{E}\varphi(\{r_{a,t}\}_{t=1}^{n(\mathcal{T}_a)}) = \mathcal{O}(\sqrt{\frac{d_x + d_\theta}{n(\mathcal{T}_a)}})$

**Remark 4.2.** An improved constant can be achieved by using weaker metric entropy bounds (namely the union bound) however this would yield a bound of order $\mathcal{O}(\sqrt{\frac{(d_x + d_\theta) \log n(\mathcal{T}_a)}{n(\mathcal{T}_a)}})$

Using the results of Lemmas 4.1–4.3, we can complete the sketch of the proof for Theorem 4.3. Lemma 4.1 says the mapping $\varphi$ is $L_p$-Lipschitz, and combining this with Assumption 4.4 implies that by Theorem 1 in (Kontorovich, 2014) we have with probability at most $\exp(\frac{-\xi^2 n(\mathcal{T}_a)}{2\epsilon^2 L_P^2 \sigma^2})$ that the maximum difference between the empirical KL divergence and the true trajectory divergence is sufficiently far from its mean. Then using Lemma 4.3 we obtain an upper bound on this expected value with the appropriate constants. For a complete proof of the theorem please refer to Appendix C.1. This theorem is useful because it indicates the empirical KL divergence derived from the MLE objective converges uniformly in probability to the true trajectory KL divergence.

### 4.3.3 Concentration of Trajectory Divergence

We can complete the proof of Theorem 4.1 using the results of Theorem 4.3 and the definition of the MLE. First, Theorem 4.3 implies that with high probability the trajectory divergence between the MLE parameters $\hat{\theta}_a, \hat{x}_{a,0}$ and true parameters $\theta_a^*, x_{a,0}^*$ is within $\mathcal{O}(\sqrt{\frac{d_x+d_\theta}{n(\mathcal{T}_a)}})$ of the empirical divergence between these two sets of parameters. Then, since $\hat{\theta}_a, \hat{x}_{a,0}$ minimize the empirical divergence and the empirical divergence of $\theta_a^*, x_{a,0}^*$ is zero, this means that the empirical divergence term is non-positive. Combining these two facts yields the concentration bound of Theorem 4.1, and the complete proof is given in Appendix C.1.

We conclude this section with an alternative statement of Theorem 4.1.

**Corollary 4.4.** For $\alpha \in (0,1)$, with probability at least $1-\alpha$ we have

$$\frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \hat{\theta}_a, \hat{x}_{a,0}) \leq B(\alpha) \sqrt{\frac{\log(1/\alpha)}{n(\mathcal{T}_a)}}. \tag{4.20}$$

Where $B(\alpha) = \frac{c_f(d_x,d_\theta)}{\sqrt{\log(1/\alpha)}} + L_p \sigma \sqrt{2}$.

This result can be obtained by making the substitution $\xi = L_p \sigma \sqrt{\frac{\log(1/\alpha)}{n(\mathcal{T}_a)}}$ into the expression in Theorem 4.1. This corollary is significant because it allows us to derive confidence bounds for our parameter estimates with regards to their trajectory divergence. Note that the term $B(\alpha)$ differs from the term that would be derived by Chernoff-Hoeffding bounds applied to i.i.d. random variables by the addition of $\frac{c_f(d_x,d_\theta)}{\sqrt{\log(1/\alpha)}}$ to the standard variance term. The reason for this addition is that since we are using MLE for our parameter estimation our estimates will be biased, and this bias must be accounted for in the confidence bounds. Though there may exist specific models where MLE can provide unbiased estimates, we will only present analysis for the more general case.

## 4.4 ROGUE Upper Confidence Bounds (ROGUE-UCB) Policy

This section develops our ROGUE-UCB policy for the ROGUE bandit model. Though several upper confidence bounds (UCB) policies have been proposed in the non-stationary setting (Garivier and Moulines, 2008, Besbes et al., 2014), these existing policies provide regret of order $\mathcal{O}(\sqrt{T \log T})$. In contrast, the ROGUE-UCB policy we construct achieves regret of order $\mathcal{O}(\log T)$, which is optimal in that it matches the lowest achievable rate for approximate policies in the stationary case.

Pseudocode for ROGUE-UCB is given in Algorithm 3, and the algorithm is written for the situation where the policy chooses actions over the course of $T$ time periods labeled $\{1, ..., T\}$. The upper confidence bounds used in this algorithm are computed using the concentration inequality from Theorem 4.1. Much like other UCB policies, for the first $|\mathcal{A}|$ time steps of the algorithm each action $a$ will be tried once. Then after this initialization, at each time step, we will first compute the MLE estimates of the parameters for each action (i.e., $(\theta_a, \hat{x}_{0,a}) \forall a \in \mathcal{A}$) and then use Theorem 4.1 to form the upper confidence bound on the value of $g(\theta_a, x_{t,a})$, which we call $g_{a,t}^{UCB}$. Our approach for forming these bounds is similar to the method first proposed by Garivier and Cappé (2011) for the KL-UCB algorithm used for stationary bandits. Here, since we know that with high probability the true parameters belong to $\mathcal{X}$ and $\Theta$, we find the largest possible value of $g(\theta_a, x_{t,a})$ within these sets. Finally, we choose the action that has the largest upper confidence bound, observe the result, and repeat the algorithm in the next time step.

The key theoretical result about the ROGUE-UCB algorithm concerns the regret $R_\Pi(T)$ of the policy computed by the ROGUE-UCB algorithm.

**Theorem 4.5.** *The expected regret $\mathbb{E}R_\Pi(T)$ for a policy $\Pi$ computed by the ROGUE-UCB algorithm is*

$$\mathbb{E}R_\Pi(T) \leq L_g \operatorname{diam}(\mathcal{X} \times \Theta) \sum_{a \in \mathcal{A}} \left( A(|\mathcal{A}|)^2 \frac{4 \log T}{\delta_a^2} + \frac{\pi^2}{3} \right). \tag{4.21}$$

*where $A(x) = B(x^{-4})$, and*

$$\begin{aligned} \delta_a = \min\{ & \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a, x_{a,0} || \theta_{a'}, x_{a',0}) : \\ & |g(h_a^t(x_{a,0}), \theta_a) - g(h_a^t(x_{a',0}), \theta_{a'})| \geq \frac{\epsilon_a}{2} \} \\ \epsilon_a = \min_{a' \in \mathcal{A} \setminus a, t} & \{ |g(\theta_a, h_a^t(x_{a,0})) - g(\theta_{a'}, h_a^t(x_{a',0}))| : \\ & g(\theta_a, h_a^t(x_{a,0})) \neq g(\theta_{a'}, h_a^t(x_{a',0})) \} \end{aligned} \tag{4.22}$$

*are finite and strictly positive constants.*

**Remark 4.3.** This corresponds to a rate of order $\mathcal{O}(\log T)$ when $\liminf_T \delta_a > 0$. In fact, $\liminf_T \delta_a > 0$ for many settings such as (with appropriate choice of model parameter values) the ROGUE GLM and ROGUE agent defined in Section 4.5.

**Algorithm 3** Reducing or Gaining Unknown Efficacy Upper Confidence Bounds (ROGUE-UCB)

```
for $t \le |\mathcal{A}|$ do
    $\pi_t = a$ such that $a$ hasn't been chosen before
end for
for $|\mathcal{A}| \le t \le T$ do
    for $a \in \mathcal{A}$ do
        Compute: $\hat{\theta}_a, x_{a,0} = \text{argmin}\{-\sum_{t\in\mathcal{T}_a} \log p(r_{a,t}|\theta_a, x_{a,t}) : x_{a,t+1} = h_a(x_{a,t}, \pi_{a,t}) \forall t \in 0, ..., T\}$
        Compute: $g^{UCB}_{a,t} = \max_{\theta_a, x_{a,0} \in \Theta\times\mathcal{X}}\{g(\theta_a, h^t_a(x_{a,0})) : \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta^*_a, x^*_{a,0}||\hat{\theta}_a, \hat{x}_{a,0}) \le A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}\}$
    end for
    Choose $\pi_t = \text{argmax}_{a\in\mathcal{A}}\, g^{UCB}_{a,t}$
end for
```

To prove Theorem 4.5, we first present two propositions. The first proposition bounds the expected regret $R_\Pi(T)$ by the number of times an action is taken while it is suboptimal.

**Proposition 4.6.** For a policy $\Pi$ calculated using the ROGUE-UCB algorithm, if $\tilde{T}_a = \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi^*_t\}$, then $\mathbb{E}R_\Pi(T) \le L_g \operatorname{diam}(\mathcal{X} \times \Theta) \sum_{a\in\mathcal{A}} \mathbb{E}\tilde{T}_a$.

For this proposition, we first use Assumption 4.3 to upper bound the value of the regret with respect to the $L_g$ and the diameter of the parameter set. Then since we are left with a finite sum of positive numbers, we can rearrange the summation term to obtain the expected number of suboptimal actions. For the detailed proof, please see Appendix C.1. Next we proceed to prove a bound on the expected number of times a suboptimal action will be chosen.

**Proposition 4.7.** For a policy $\Pi$ calculated using the ROGUE-UCB algorithm, we have that $\mathbb{E}\tilde{T}_a \le A(|\mathcal{A}|)^2 \frac{4\log T}{\delta_a^2} + \frac{\pi^2}{3}$, where $A(t) = B(t^{-4})$, $\delta_a = \min\{\frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a, x_{a,0}||\theta_{a'}, x_{a',0}) : |g(h^t_a(x_{a,0}), \theta_a) - g(h^t_a(x_{a',0}), \theta_{a'})| \ge \frac{\epsilon_a}{2}\}$, and $\epsilon_a = \min_{a'\in\mathcal{A}\backslash a,t}\{|g(\theta_a, h^t_a(x_{a,0})) - g(\theta_a, h^t_a(x_{a,0}))|: g(\theta_a, h^t_a(x_{a,0})) \neq g(\theta_a, h^t_a(x_{a,0}))\}$.

To prove this proposition, we proceed in a manner similar to the structure first proposed by Auer et al. (2002a). We must show that if an action is chosen at a time when it is suboptimal, then this implies that either we have not properly estimated its parameters (i.e., have not explored enough) or the true values of the parameters $x_{a,0}, \theta_a$ or $x_{\pi^*_t,0}, \theta_{\pi^*_t}$ are not contained inside their confidence bounds. Using these facts, we use Theorem 4.1 to show that the probability that all of these events occurring simultaneously is bounded, and then upper bound the expected number of times these events can occur. Combining the results of Propositions 4.6 and 4.7, we thus prove the desired result of Theorem 4.5. The full proofs of Proposition 4.7 and Theorem 4.5 are provided in Appendix C.1.

## 4.5 Numerical Experiments

In this section, we perform two numerical experiments where the policy computed by the ROGUE-UCB algorithm is compared against the policy computed by other non-stationary bandit algorithms. The first experiment considers the ROGUE GLM described in Section 4.2.4, and specifically looks at the logistic regression instantiation of ROGUE GLM. We use synthetically generated data for this first experiment. Next, we perform an experiment in the context of healthcare-adherence improving interventions to increase physical activity, which can be modeled using the ROGUE agent from Section 4.2.4. Using real world data from the mDPP trial (Fukuoka et al., 2015), we show how ROGUE-UCB can be implemented to personalize messages for participants in this intervention. All experiments in this section were run using Python 3.5.2 and Anaconda on a laptop computer with a 2.4GHz processor and 16GB RAM.

### 4.5.1 Tuned ROGUE-UCB

As has been noted for other UCB policies (Auer et al., 2002a, Garivier and Moulines, 2008, Bouneffouf and Féraud, 2016), the high probability bounds derived theoretically for these methods are often too conservative. While the $\mathcal{O}(\sqrt{\frac{\log t}{n(\mathcal{T}_a)}})$ is a tight rate, the term $A(t)$ is too conservative. Drawing inspiration from Auer et al. (2002a) who used asymptotic bounds for Tuned UCB, we similarly construct a variant of our algorithm: This variant is described in Algorithm 4 and called Tuned ROGUE-UCB. Using the results of Shapiro (1993), we note that if the MLE $\hat{\theta}_a, \hat{x}_{a,0}$ are in the interior of the feasible region and are consistent, then they are asymptotically normally distributed with a variance equal to their Fisher information. Using these results and the delta method (Qu and Keener, 2011), we can derive the quantity $\mathcal{S}_{a,\pi_1^T}(\theta_a, x_{a,0}||\hat{\theta}_a, \hat{x}_{a,0}) = \frac{1}{n(\mathcal{T}_a)^2}\nabla_{\theta',x'}D_{a,\pi_1^T}(\theta_a, x_{a,0}||\theta', x')^T\mathcal{I}_{\{r_t\}_{t\in\mathcal{T}_a}}(\theta', x')^{-1}\nabla_{\theta',x'}D_{a,\pi_1^T}(\theta_a, x_{a,0}||\theta', x')|_{\theta',x'=\hat{\theta}_a,\hat{x}_{a,0}}$, which is the asymptotic variance of the average trajectory KL-Divergence. Here, $\eta$ is a constant that corresponds to the maximum value of the KL-divergence; $\mathcal{I}_{\{r_t\}_{t\in\mathcal{T}_a}}(\theta', x')$ represents the observed trajectory Fisher information, which can be calculated as $\mathcal{I}_{\{r_t\}_{t\in\mathcal{T}_a}}(\theta', x') = \sum_{t\in\mathcal{T}_a}\mathcal{I}_{r_t}(\theta', x')$, due to Assumption 4.1. As an implementation note, if the empirical information matrix is singular, then the Moore-Penrose pseudoinverse should be used to achieve similar asymptotic results (Hero et al., 1997). Note that although these asymptotic bounds work well in practice, they are not high probability bounds and do not provide the same theoretical guarantees as the ROGUE-UCB algorithm. A full analysis of regret for Tuned ROGUE-UCB is beyond the scope of this work. Instead, we only consider empirical analysis of this algorithm to show its strong performance.

**Algorithm 4** Tuned ROGUE-UCB

```
for $t \leq |\mathcal{A}|$ do
    $\pi_t = a$ such that $a$ hasn't been chosen before
end for
for $|\mathcal{A}| \leq t \leq T$ do
    for $a \in \mathcal{A}$ do
        Compute: $\hat{\theta}_a, x_{a,0} = \text{argmin}\{-\sum_{t \in \mathcal{T}_a} \log p(r_{a,t}|\theta_a, x_{a,t}) : x_{a,t+1} = h_a(x_{a,t}, \pi_{a,t}) \forall t \in 0, ..., T\}$
        Compute: $g_{a,t}^{UCB} = \max_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \left\{ g(\theta_a, h_a^t(x_{a,0})) : \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a, x_{a,0}||\hat{\theta}_a, \hat{x}_{a,0}) \leq \sqrt{\min\{\frac{\eta}{4}, \mathcal{S}_{a,\pi_1^T}(\theta_a, x_{a,0}||\hat{\theta}_a, \hat{x}_{a,0})\} \frac{\log(t)}{n(\mathcal{T}_a)}} \right\}$
    end for
    Choose $\pi_t = \text{argmax}_{a \in \mathcal{A}} g_{a,t}^{UCB}$
end for
```

### 4.5.2 Experimental Design

We examined two settings for our experiments, which correspond to the instantiations of ROGUE bandits presented in Sections 4.2.4 and 4.2.4. For each of the scenarios, we compared the Tuned ROGUE-UCB algorithm to policies determined by five alternative methods. For each scenario, we present two result metrics: cumulative regret of each algorithm in that scenario and the average reward to date of the algorithm. While these two measures are related, a key difference is that in the non-stationary setting suboptimal actions may not have a significantly lower expected reward than the optimal action at all time periods. Hence, while an algorithm may incur a significant amount of regret it could still achieve a high amount of reward. The five alternative algorithms we used for comparison are as follows:

1. **Pure Exploration:** First, we considered a completely random, or "pure exploration" algorithm, which chooses an action uniformly at random from the set of available actions.

2. **Stationary Upper Confidence Bound (UCB1):** Next, we considered the UCB1 algorithm (Auer et al., 2002a), which is designed for stationary bandits.

This approach uses the sample average as an estimate of the expected reward of each action and utilizes a padding upper confidence bound term derived from Hoeffding's bound. In our experiments, we implemented Tuned UCB1 (Auer et al., 2002a), which replaces the theoretical constants by the asymptotic variance of the sample average and a small constant that corresponds to the maximum variance of a Bernoulli random variable (since the rewards are bounded between 0 and 1).

3. **Discounted Upper Confidence Bounds (D-UCB):** D-UCB is an upper confidence bound approach designed for non-stationary systems. It utilizes an exponentially weighted average of the reward observations to estimate the expected reward at the current time period and a square root padding function to provide upper confidence bounds (Garivier and Moulines, 2008). The weighted average is constructed with a positive discount factor that decreases the influence of older observations on the reward estimate to zero as time goes on. We implemented this algorithm with its optimal theoretical parameters, as described in Garivier and Moulines (2008).

4. **Sliding Window Upper Confidence Bounds (SW-UCB):** The next approach we considered is the SW-UCB approach. This algorithm considers a fixed window size of how many action choices to "keep in memory", and computes the estimate of the expected action rewards as the average of these choices (Garivier and Moulines, 2008). We implemented this algorithm with its optimal theoretical parameters as proposed by Garivier and Moulines (2008).

5. **Exploration and Exploitation with Exponential Weights (EXP3):** The last bandit algorithm we considered in our experiments is the EXP3 algorithm. Essentially, EXP3 is a modification of the exponential weights algorithm used in online optimization to the bandit setting where not all action rewards are observed (Auer et al., 2002b). Though EXP3 is designed for stationary bandits, unlike UCB approaches that assume a stochastic setting, it is meant for adversarial bandits, which makes it potentially robust to non-stationarity. The particular variant of EXP3 we utilized is EXP3.S proposed by Auer et al. (2002b), which is designed for arbitrary reward sequences, using the theoretically optimal parameters as proposed by the authors.

### 4.5.3 ROGUE Logistic Regression

For this experiment, we consider the logistic regression instantiation of the ROGUE GLM presented in Section 4.2.4. Our setup includes two actions whose rewards $r_{t,a}$ are Bernoulli with a logistic link function of the form $g(x, \theta) = \frac{1}{1+\exp(-a\theta-bx)}$. The initial parameters and dynamics matrices for each of the actions are presented in Table 4.1. Here, the sets $\mathcal{X}$ and $\Theta$ were set to $[0, 1]$. Action 0 has a significant portion of its reward dependent on the time varying state $x_t$, and recovers its reward slowly but also decreases slowly. On the other hand, Action 1 has more of its expectation dependent on the stationary component $\theta$, but it expectation decreases faster than that of Action 0.

| Action | $x_0$ | $\theta$ | $A$ | $B$ | $K$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|---|
| 0 | 0.1 | 0.5 | 0.6 | -1.0 | 0.5 | 0.4 | 0.6 |
| 1 | 0.3 | 0.7 | 0.7 | -1.2 | 0.5 | 0.7 | 0.3 |

Table 4.1: Experimental parameters for each action for the logistic ROGUE GLM simulation

The experiments were run for 20,000 action choices and replicated 30 times for each of the candidate algorithms. Figure 4.1 shows the cumulative regret accrued by each of the algorithms averaged across the replicates, and Figure 4.2 shows the average reward per action for each algorithm averaged across the replicates. As expected in these experiments, the UCB1 algorithm achieves linear regret since it assumes a stationary model and thus converges to a single action, which causes a large gap between the expectations of the two actions. Interestingly, SW-UCB and D-UCB also perform worse than random choices. A key note here is that D-UCB and SW-UCB assume that action rewards do not change frequently and are independent of the choices. However, D-UCB outperforms SW-UCB since the weighted average contains more information about the trajectory of the expected reward of each action while data from earlier choices are removed from the estimates in the sliding window. EXP3 and random action selection perform approximately the same in terms of both regret and expected reward. This is unsurprising because the weighting scheme in EXP3 emphasizes the rewards of the past action states as opposed to current action states. In terms of both regret and reward, Tuned ROGUE-UCB substantially outperforms the other approaches. While the other approaches seem to obtain linear regret, ROGUE-UCB does in fact have regret on the order of $\mathcal{O}(\log T)$ in this experiment.

### 4.5.4 Healthcare-Adherence Improving Intervention for Increasing Physical Activity

Next, we consider an experiment using real world data from the mobile diabetes prevention program (mDPP) (Fukuoka et al., 2015). This was a randomized control trial (RCT) that was conducted to evaluate the efficacy of a 5 month mobile phone based

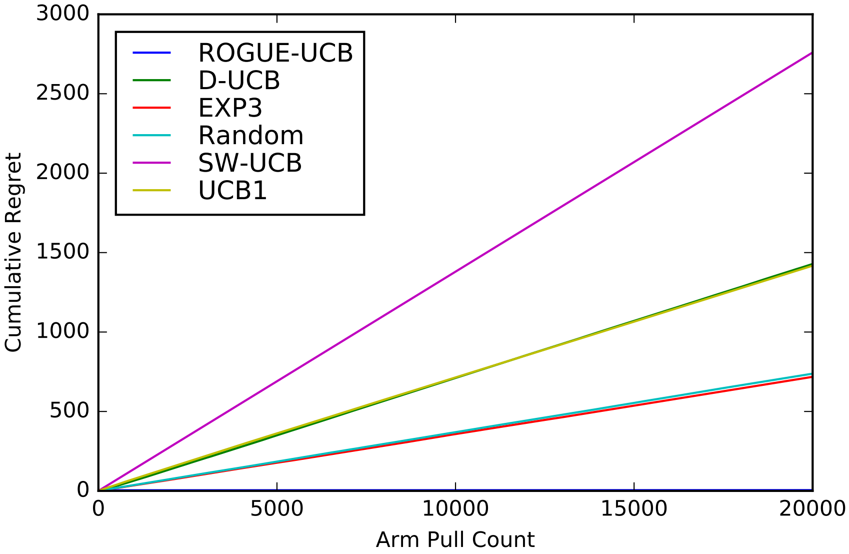


Figure 4.1: Comparison of cumulative regret between the different bandit algorithms for the logistic ROGUE GLM.

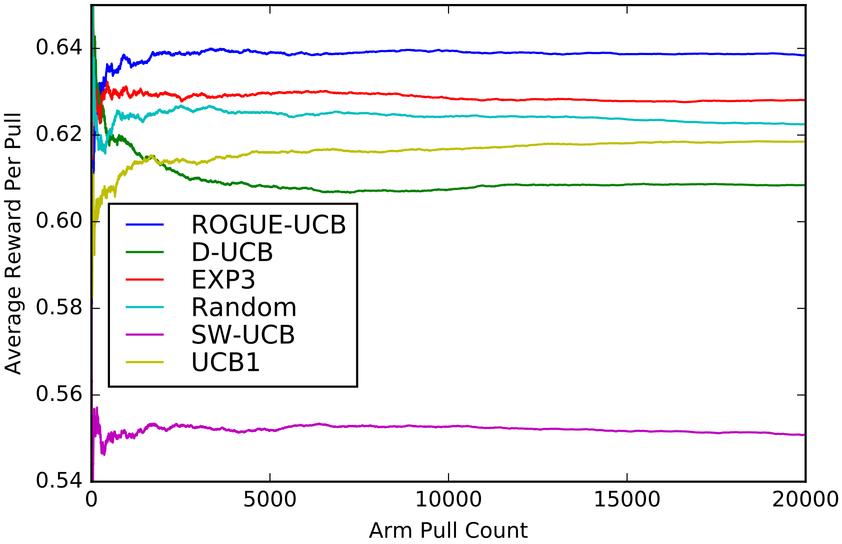


Figure 4.2: Comparison of average reward to date between the different bandit algorithms for the logistic ROGUE GLM.

weight loss program among overweight and obese adults at risk for developing type 2 diabetes and was adapted from the Diabetes Prevention Program (DPP) (Diabetes Prevention Program Research Group, 2002, 2009). Sixty one overweight/obese adults were randomized into an active control group that only received an accelerometer (n=31) or a treatment group that received the mDPP mobile app plus the accelerometer and clinical office visits (n=30). Changes in primary and secondary outcomes for the trial were clinically and statistically significant. The treatment group lost an average of 6.2 ± 5.9 kg (-6.8% ± 5.7%) between baseline and the 5 month follow up while the control group gained 0.3 ± 3.0 kg (0.3% ± 5.7 %) ($p < 0.001$). The treatment group's steps per day increased by 2551 ± 4712 compared to the control group's decrease of 734 ± 3308 steps per day ($p < 0.001$). Additional details on demographics and other treatment parameters are available in (Fukuoka et al., 2015).

One key feature of the mDPP application was the ability for the clinicians to send daily messages to the participants to encourage that they adhere to the intervention and maintain a sufficiently increased activity level. Broadly speaking, there were 5 different message categories that the clinicians could choose to send to the patients. These categories are self-efficacy/confidence, motivation/belief/attitude, knowledge, behavior reinforcement, and social support. Each day the experimental group would receive a preprogrammed message from one of these categories, and all participants received the same messages each day. For our simulations, we used the data of what messages were sent to what participants, as well as their daily step counts.

### Patient Model

For our experiment, we used a behavioral analytics model of patient behavior first proposed by Aswani et al. (2016). Here, each patient is assumed to be a utility maximizing agent who chooses how many steps to take each day based on previous behavior and the intervention implemented. We defined each of the different message categories be one of the actions of the bandit, which forms a ROGUE agent model as described in Section 4.2.4. Using the notation of Section 4.2.4, let $c_t$ be a sequence of i.i.d. Laplace random variables with mean zero and shape parameter $\theta$. This means $\sigma^2(\theta) = 2\theta^2$. After normalizing the step counts to be in $[0, 1]$ (where 1 is equal 14,000 steps), we can then write the reward distribution of a particular message type $a$ as $p(r_t|\{x_{a,t}, \theta_a\}) = \frac{1}{2}\exp(\frac{-x_{a,t}}{\theta_a})\delta(r_t) + \frac{1}{2}\exp(\frac{x_{a,t}-1}{\theta_a})\delta(1-r_t) + \frac{1}{2\theta_a}\exp(\frac{-|r_t-x_t|}{\theta_a})\mathbf{1}[r_t \in (0, 1)]$, where the state $x_{a,t} \in [0, 1]$ and $\theta_a \in [\epsilon, 1]$ for a small $\epsilon > 0$. This results in a reward function $g(x, \theta) = x + \frac{\theta}{2}(\exp(\frac{-x}{\theta}) - \exp(\frac{x-1}{\theta}))$. Using Laplace noise has the advantage of allowing commercial mixed integer programming solvers to be used for offline parameter estimation by solving inverse optimization problems (Aswani et al., 2015, Aswani, 2017, Mintz et al., 2017a). Using this MILP reformulation and behavioral models, we estimated the respective trajectory parameters for each message group and each patient of the treatment group for which we had data. These initial parameters were found using the Gurobi Solver in Python (Gurobi Optimization, 2015).

### Simulation Results

This simulation was conducted using the mDPP data described above. Each experiment consisted of 1,000 action choices, which would correspond to about two years of a message based physical activity intervention, and 10 replicates of the simulation were conducted per patient and algorithm. The results in Figures 4.3 and 4.4 represent averages across all patients and replicates. Since we are using real data, the interpretation of the y-axis of each of the plots corresponds to number of steps in units of 1,000 steps, and the x-axis corresponds to the day of the intervention.

ROGUE-UCB outperforms all other algorithms both in terms of regret and average reward. In terms of regret, ROGUE-UCB is the only algorithm that obtains logarithmic regret. While D-UCB is the only other algorithm that can outperform pure exploration, it only obtains linear regret. In terms of average reward, ROGUE-UCB and D-UCB are the only two algorithms that outperform pure exploration. Interpreting these results in the healthcare context of this intervention, we find that the improved predictive model and use of MLE estimates within our ROGUE-UCB algorithm results in an increase of 1,000 steps a day (approximately a half-mile more of walking per day) relative to the next best algorithm, which is a significant increase in activity.

## 4.6 Conclusion

In this chapter, we defined a new class of non-stationary bandit models where the specific actions chosen influence the reward distributions of each action in subsequent time periods through a specific model. We conducted a finite sample analysis of the MLE estimates in this setting, and showed how these concentration bounds can be used to create a ROGUE-UCB algorithm that provides a policy for these bandit models. Our theoretical results show that in expectation ROGUE-UCB achieves logarithmic regret. This is a substantial improvement over model-free algorithms, which can only achieve a square-root regret. We then showed through simulations using real and artificial data, that with minor modification, the ROGUE-UCB algorithm significantly outperforms state of the art bandit algorithms both in terms of cumulative regret and average reward. These results suggest that ROGUE bandits have strong potential for personalizing health care interventions, and in particular for healthcare-adherence improving interventions.

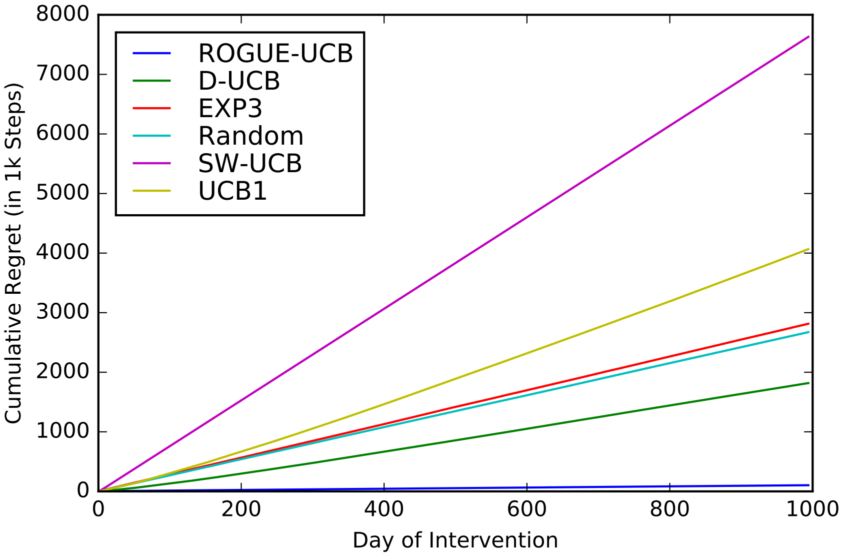


Figure 4.3: Comparison of cumulative regret between the different bandit algorithms for the healthcare-adherence improving intervention.

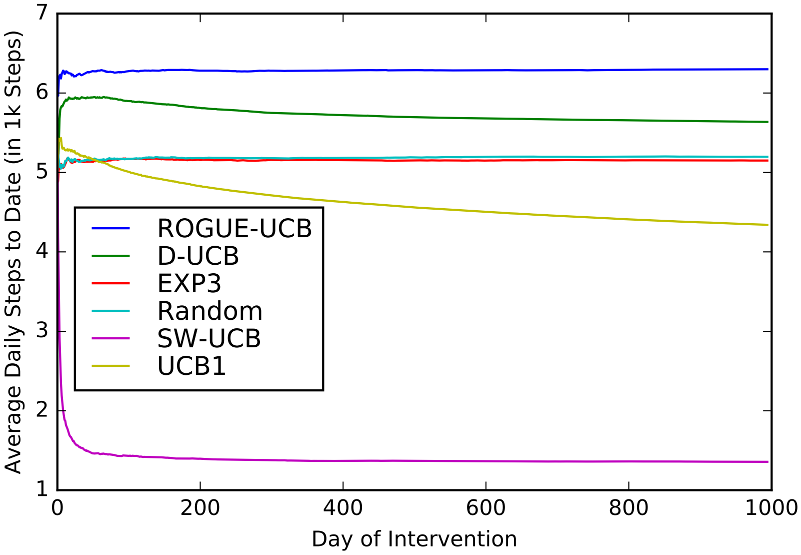


Figure 4.4: Comparison of average reward to date between the different bandit algorithms for the healthcare-adherence improving intervention.

# Chapter 5

# Conclusion

In this thesis we described a new precision analytics framework that can be used to design and optimize personalized systems. While for the most part the technical implications and future work of the methodologies was described at the end of each of the chapters, these models will need to be extended to accommodate additional engineering specification.

For instance, a natural extension of the methodologies in this thesis would include incorporating notions of risk into the policy calculation and parameter estimation. While we discussed and developed interpretable models which account for human behavior and used them for optimization, these were primarily designed with low risk scenarios in mind (such as fitness tracking or advertisement targeting) where the cost of making a single bad prediction is minimal. However, if we consider other riskier scenarios such as personalized diagnosis where the cost of a wrong policy may be life or death, we need to extend the models to provide additional safety guarantees while still being effective.

In addition, while we only considered numerical or categorical data in our analysis another extension of these methods could include incorporating new forms of unorganized and qualitative data (e.g. speech, text, images etc.) to improve prediction accuracy and intervention efficacy. There exist many examples of precision systems which collect qualitative data in the form of text and images as feedback from participants. For instance, in the case of diet tracking participants are often asked to provide images of their meals and have text or in person conversations with their clinician which are later transcribed. State of the art methods for processing this type of data often involve deep learning models which are challenging to incorporate into reinforcement learning and precision analytics settings and thus require additional anlysis to be properly incorporated into a precision analytics framework.

# Bibliography

Acharya, S., Elci, O., Sereika, S., Music, E., Styn, M., Turk, M., and Burke, L. (2009). Adherence to a behavioral weight loss treatment program enhances weight loss and improvements in biomarkers. *Patient Preference and Adherence*, 3:151–160.

Adlakha, S. and Johari, R. (2013). Mean field equilibrium in dynamic games with strategic complementarities. *Operations Research*, 61(4):971–989.

Afèche, P. and Ata, B. (2013). Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management*, 15(2):292–304.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135.

Agrawal, S., Zizhuo, Y., and Ye, Y. (2014). A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890.

Ahuja, R. and Orlin, J. (2001). Inverse optimization. *Operations Research*, 49(5):771–783.

Ajzen, I. and Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall.

Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. *IEEE TAC*, 32(11):977–982.

Andersen, A. R., Nielsen, B. F., and Reinhardt, L. B. (2017). Optimization of hospital ward resources with patient relocation using markov chain modeling. *European Journal of Operational Research*, 260(3):1152–1163.

Astrom, K. and Wittenmark, B. (1995). *Adaptive Control*. Addison–Wesley.

Aswani, A. (2017). Statistics with set-valued functions: Applications to inverse approximate optimization. *arXiv preprint*.

Aswani, A., Gonzalez, H., Sastry, S., and Tomlin, C. (2013). Provably safe and robust learning–based model predictive control. *Automatica*, 49(5):1216–1226.

Aswani, A., Kaminsky, P., Mintz, Y., Flowers, E., and Fukuoka, Y. (2016). Behavioral modeling in weight loss interventions. Available at SSRN: https://ssrn.com/abstract=2838443.

Aswani, A., Shen, Z.-J., and Siddiq, A. (2018). Inverse optimization with noisy data. *Operations Research*. To appear.

Aswani, A., Shen, Z.-J. M., and Siddiq, A. (2015). Inverse optimization with noisy data. *arXiv preprint*.

Aswani, A. and Tomlin, C. (2009). Monotone piecewise affine systems. *IEEE Transactions on Automatic Control*, 54(8):1913–1918.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Ayer, T., Alagoz, O., and Stout, N. (2012). A POMDP approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Ayer, T., Alagoz, O., Stout, N., and Burnside, E. (2015). Heterogeneity in women's adherence and its role on optimal breast cancer screening policies. *Management Science*.

Azar, K., Lesser, L., Laing, B., Stephens, J., Aurora, M., Burke, L., and Palaniappan, L. (2013). Mobile applications for weight management: theory–based content analysis. *American Journal of Preventive Medicine*, 45(5):583–589.

Ban, G. (2015). The data-driven (s, s) policy: Why you can have confidence in censored demand data. *Available at SSRN*.

Ban, G. and Rudin, C. (2016). The big data newsvendor: Practical insights from machine learning.

Ban, G.-Y. and Rudin, C. (2014). The big data newsvendor: Practical insights from machine learning. *Available on SSRN `https://ssrn.com/abstract=2559116`.*

Bandura, A. (1998). Health promotion from the perspective of social cognitive theory. *Psychology and Health*, 13(4):623–649.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52(1):1–26.

Bartlett, P. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bastani, H. and Bayati, M. (2015a). Online decision-making with high-dimensional covariates. *Available at SSRN 2661896*.

Bastani, H. and Bayati, M. (2015b). Online decision-making with high-dimensional covariates.

Bender, M., Choi, J., Arai, S., Paul, S., Gonzalez, P., and Fukuoka, Y. (2014). Digital technology ownership, usage, and factors predicting downloading health apps among caucasian, filipino, korean, and latino americans: the digital link to health survey. *JMIR Mhealth and Uhealth*, 2(4):e43.

Bertsimas, D. and Goyal, V. (2012). On the power and limitations of affine policies in two-stage adaptive optimization. *Mathematical programming*, 134(2):491–531.

Bertsimas, D., Gupta, V., and Paschalidis, I. (2014). Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, pages 1–39.

Bertsimas, D. and Mersereau, A. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135.

Bertsimas, D. and Nino-Mora, J. (1994). Restless bandits, linear programming relaxations and a primal-dual heuristic.

Bertsimas, D. and Niño-Mora, J. (2000). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90.

Bertsimas, D. and O'Hair, A. (2013). Personalized diabetes management: A robust optimization approach. Submitted.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *arXiv preprint arXiv:1405.3316*.

Besbes, O., Gur, Y., and Zeevi, A. (2015). Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244.

Bickel, P. and Doksum, K. (2006). *Mathematical Statistics: Basic Ideas And Selected Topics*, volume 1. Pearson Prentice Hall, 2nd edition.

Bitmead, R. (1984). Persistence of excitation conditions and the convergence of adaptive schemes. *IEEE Transactions on Information Theory*, 30(2):183–191.

Blanchet, J., Gallego, G., and Goyal, V. (2013). A markov chain approximation to choice modeling. In *EC*, pages 103–104. Citeseer.

Boukhtouta, A., Berger, J., Powell, W., and George, A. (2011). An adaptive-learning framework for semi-cooperative multi-agent coordination. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 324–331. IEEE.

Bouneffouf, D. and Féraud, R. (2016). Multi-armed bandit problem with known trend. *Neurocomputing*, 205:16–21.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Brock, D. and Wartman, S. (1990). When competent patients make irrational choices. *New England Journal of Medicine*, 322(22):1595–1599.

Callier, F. and Desoer, C. (1994). *Linear System Theory*. Springer Texts in Electrical Engineering. Springer New York.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292.

Cawley, J. (2004). An economic framework for understanding physical activity and eating behaviors. *American Journal of Preventive Medicine*, 27(3, Supplement):117–125.

Chang, H., Fu, M., Hu, J., and Marcus, S. (2005). An adaptive sampling algorithm for solving markov decision processes. *Operations Research*, 53(1):126–139.

Chapman, G. and Elstein, A. (1995). Valuing the future: temporal discounting of health and money. *Medical Decision Making*, 15(4):373–386.

Chen, C., Haddad, D., Selsky, J., Hoffman, J., Kravitz, R., Estrin, D., and Sim, I. (2012). Making sense of mobile health data: An open architecture to improve individual- and population-level health. *Journal of Medical Internet Research*, 14(4):e112.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework, results and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159.

Committee, P. A. G. A. et al. (2008). Physical activity guidelines for americans. *Washington, DC: US Department of Health and Human Services*, pages 15–34.

Conner, M. and Norman, P. (1996). *Predicting health behaviour: research and practice with social cognition models*. Open University Press.

Corbett, C. and Tang, C. (1999). Designing supply contracts: Contract type and information asymmetry. In *Quantitative models for supply chain management*, pages 269–297. Springer.

Costa, L. and Kariniotakis, G. (2007). A stochastic dynamic programming model for optimal use of local energy resources in a market environment. In *Power Tech, 2007 IEEE Lausanne*, pages 449–454. IEEE.

Craig, J., Hsu, P., and Sastry, S. (1987). Adaptive control of mechanical manipulators. *The International Journal of Robotics Research*, 6(2):16–28.

Darby, S. (2010). Smart metering: What potential for householder engagement? *Building Research and Information*, 38(5):442–457.

Dasgupta, S. and Huang, Y.-F. (1987). Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise. *IEEE Transactions on information theory*, 33(3):383–392.

Dempe, S. (2002). *Foundations of Bilevel Programming*. Springer.

DeNegre, S. and Ralphs, T. (2009). A branch-and-cut algorithm for bilevel integer programming. In *Proceedings of the Eleventh INFORMS Computing Society Meeting*, pages 65–78.

Deng, R., Yang, Z., Chow, M. Y., and Chen, J. (2015). A survey on demand response in smart grids: Mathematical models and approaches. *IEEE Transactions on Industrial Informatics*, 11(3):570–582.

Denoyel, V., Alfandari, L., and Thiele, A. (2017). Optimizing healthcare network design under reference pricing and parameter uncertainty. *European Journal of Operational Research*, 263(3):996–1006.

Deo, S., Jiang, T., Iravani, S., Smilowitz, K., and Samuelson, S. (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research*, 61(6):1277–1294.

Desai, V. (1992). Marketing-production decisions under independent and integrated channel structure. *Annals of Operations Research*, 34(1):275–306.

Desai, V. (1996). Interactions between members of a marketing-production channel under seasonal demand. *European Journal of Operational Research*, 90(1):115–141.

Diabetes Prevention Program Research Group (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine*, 346(6):393–403.

Diabetes Prevention Program Research Group (2003). Costs associated with the primary prevention of type 2 diabetes mellitus in the diabetes prevention program. *Diabetes Care*, 26(1):36–47.

Diabetes Prevention Program Research Group (2009). 10–year follow–up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *Lancet*, 374(9720):1677–1686.

Doucet, E., St-Pierre, S., Alméras, N., Després, J.-P., Bouchard, C., and Tremblay, A. (2001). Evidence for the existence of adaptive thermogenesis during weight loss. *British Journal of Nutrition*, 85(6):715–723.

Eliashberg, J. and Steinberg, R. (1987). Marketing-production decisions in an industrial channel of distribution. *Management Science*, 33(8):981–1000.

Engineer, F., Keskinocak, P., and Pickering, L. (2009). Or practice – catch-up scheduling for childhood vaccination. *Operations Research*, 57(6):1307–1319.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical science*, pages 254–272.

Feighery, E., Ribisl, K., Clark, P., and Haladjian, H. (2003). How tobacco companies ensure prime placement of their advertising and products in stores: interviews with retailers about tobacco company incentive programmes. *Tobacco Control*, 12(2):184–188.

Fetta, A., Harper, P., Knight, V., and Williams, J. (2018). Predicting adolescent social networks to stop smoking in secondary schools. *European Journal of Operational Research*, 265(1):263–276.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Flegal, K., Carroll, M., Kit, B., and Ogden, C. (2012). Prevalence of obesity and trends in the distribution of body mass index among U.S. adults, 1999–2010. *Journal of the American Medical Association*, 307(5):491–497.

Flores Mateo, G., Granado-Font, E., Ferré-Grau, C., and na Carreras, X. M. (2015). Mobile phone apps to promote weight loss and increase physical activity: A systematic review and meta-analysis. *J Med Internet Res*, 17(11):e253.

Fortescue, T., Kershenbaum, L. S., and Ydstie, B. E. (1981). Implementation of self-tuning regulators with variable forgetting factors. *Automatica*, 17(6):831–835.

Frazier, P. and Wang, J. (2016). Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75. Springer.

Friedenreich, C., Neilson, H., and Lynch, B. (2010). State of the epidemiological evidence on physical activity and cancer prevention. *European journal of cancer*, 46(14):2593–2604.

Fudenberg, D., Levine, D., and Maskin, E. (1994). The folk theorem with imperfect public information. *Econometrica*, 62(5):997–1039.

Fukuoka, Y., Gay, C., Joiner, K., and Vittinghoff, E. (2015). A novel diabetes prevention intervention using a mobile app: A randomized controlled trial with overweight adults at risk. *American Journal of Preventive Medicine*, 49(2):223–237.

Fukuoka, Y., Joiner, K., and Vittinghoff, E. (2014). A novel approach to reduce risk of type 2 diabetes – pilot randomized controlled clinical trial. *Diabetes Care*. In progress.

Fukuoka, Y., Komatsu, J., Suarez, L., Vittinghoff, E., Haskell, W., Noorishad, T., and Pham, K. (2011). The mPED randomized controlled clinical trial: applying mobile persuasive technologies to increase physical activity in sedentary women protocol. *BMC Public Health*, 11(933).

Gafni, A. (1990). Correspondence to “competent patients and irrational choices”. *New England Journal of Medicine*, 323(19):1353–1355.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376.

Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

George, A. and Powell, W. (2007). An adaptive-learning framework for semi-cooperative multi-agent coordination.

Ghose, A. and Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177.

Glover, F. (1975). Improved linear integer programming formulations of nonlinear integer problems. *Management Science*, 22(4):455–460.

Goldfarb, A. and Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404.

Goldfarb, A. and Tucker, C. (2014). Standardization and the effectiveness of online advertising. *Management Science*, 61(11):2707–2719.

Gomes, N., Merugu, D., O'Brien, G., Mandayam, C., Yue, J. S., Atikoglu, B., Albert, A., Fukumoto, N., Liu, H., Prabhakar, B., and Wischik, D. (2012). Steptacular: An incentive mechanism for promoting wellness. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pages 1–06.

Gonzalez, V., Goeppinger, J., and Lorig, K. (1990). Four psychosocial theories and their application to patient education and clinical practice. *Arthritis Care and Research*, 3(3):132–143.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Guha, S., Munagala, K., and Shi, P. (2010). Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):3.

Gupta, D. and Wang, L. (2008). Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56:576.

Gurobi Optimization, I. (2015). Gurobi optimizer reference manual.

Gutierrez, G. and He, X. (2011). Life-cycle channel coordination issues in launching an innovative durable product. *Production and Operations Management*, 20(2):268–279.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., and Sebag, M. (2006). Multi-armed bandit, dynamic environments and meta-bandits.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag, 2nd edition.

He, X., Prasad, A., Sethi, S., and Gutierrez, G. (2007). A survey of stackelberg differential game models in supply and marketing channels. *Journal of Systems Science and Systems Engineering*, 16(4):385–413.

He, X. and Sethi, S. (2008). Dynamic slotting and pricing decisions in a durable product supply chain. *Journal of Optimization Theory and Applications*, 137(2):363–379.

Helm, J. E., Lavieri, M. S., Oyen, M. P. V., Stein, J. D., and Musch, D. C. (2015). Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research*.

Hero, A., Usman, M., Sauve, A., and Fessler, J. (1997). Recursive algorithms for computing the cramer-rao bound. *IEEE transactions on signal processing*, 45(3):803–807.

Heuberger, C. (2004). Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization*, 8(3):329–361.

Hill, J., Wyatt, H., Reed, G., and Peters, J. (2003). Obesity and the environment: Where do we go from here? *Science*, 299(5608):853–855.

Iyer, K., Johari, R., and Moallemi, C. (2014). Information aggregation and allocative efficiency in smooth markets. *Management Science*, 60(10):2509–2524.

Iyer, K., Johari, R., and Sundararajan, M. (2011). Mean field equilibria of dynamic auctions with learning. *ACM SIGecom Exchanges*, 10(3):10–14.

James, J. M. and Bard, J. (1990). The mixed integer linear bilevel programming problem. *Oper. Res.*, 38(5):911–921.

Janz, N. and Becker, M. (1984). The health belief model: A decade later. *Health Education & Behavior*, 11(1):1–47.

Jiang, L. and Low, S. (2011). Real-time demand response with uncertain renewable energy in smart grid. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 1334–1341. IEEE.

Johari, R., Pekelis, L., and Walsh, D. (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Joos, S. and Hickam, D. (1990). How health professionals influence health behavior: patient provider interaction and health care outcomes. In *Health behavior and health education: theory, research and practice*, pages 216–241. Jossey-Bass.

Kanfer, F. (1975). Self-management methods. In *Helping People Change*, pages 309–316. Pergamon.

Kaut, M. and Wallace, S. (2003). Evaluation of scenario-generation methods for stochastic programming.

Keshavarz, A., Wang, Y., and Boyd, S. (2011). Imputing a convex objective function. In *IEEE Multi-Conference on Systems and Control*, pages 613–619.

Kim, M. and Lim, A. (2015). Robust multiarmed bandit problems. *MS*, 62(1):264–285.

Kim, T. and Poor, H. (2011). Scheduling power consumption with price uncertainty. *IEEE Transactions on Smart Grid*, 2(3):519–527.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM.

Kogan, K. and Tapiero, C. (2007). *Supply chain games: operations management and risk valuation*, volume 113. Springer Science & Business Media.

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *ICML*, pages 28–36.

Koolen, W., Malek, A., and Bartlett, P. (2014). Efficient minimax strategies for square loss games. In *Advances in Neural Information Processing Systems*, pages 3230–3238.

Koolen, W., Malek, A., Bartlett, P., and Abbasi, Y. (2015). Minimax time series prediction. In *Advances in Neural Information Processing Systems*, pages 2557–2565.

Kucukyazici, B., Verter, V., and Mayo, N. (2011). An analytical framework for designing community-based care for chronic diseases. *POMS*, 20:474.

Laffont, J.-J. and Martimort, D. (2002). *The Theory of Incentives: The Principal-Agent Model*. Princeton.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lee, H. and Whang, S. (2000). Information sharing in a supply chain. *International Journal of Manufacturing Technology and Management*, 1(1):79–93.

Lee, M., Aslam, O., Foster, B., Kathan, D., and Young, C. (2014). Assessment of demand response and advanced metering. *Federal Energy Regulatory Commission, Staff Rep.*

Leff, H., Dada, M., and Graves, S. (1986). An lp planning model for a mental health community support system. *Management Science*, 32:139.

Lehmann, E. and Romano, J. (2006). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer.

Leung, S.-H. and So, C. (2005). Gradient-based variable forgetting factor rls algorithm in time-varying environments. *IEEE Transactions on Signal Processing*, 53(8):3141–3150.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *arXiv preprint arXiv:1702.07274*.

Lewis, B., Napolitano, M., Buman, M., Williams, D., and Nigg, C. (2017). Future directions in physical activity intervention research: expanding our focus to sedentary behaviors, technology, and dissemination. *Journal of behavioral medicine*, 40(1):112–126.

Li, N., Chen, L., and Low, S. H. (2011). Optimal demand response based on utility maximization in power networks. In *2011 IEEE power and energy society general meeting*, pages 1–8. IEEE.

Li, P., Lim, A. E. B., and Shanthikumar, J. G. (2012). Decentralized control of a stochastic multi-agent queueing system. *IEEE Transactions on Automatic Control*, 57(11):2762–2777.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. In *Proc. 4th Berkeley Symp. on Math. Stat. and Prob*, pages 453–468.

Liu, K. and Zhao, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567.

Liu, N., Ziya, S., and Kulkarni, V. (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *M&SOM*, 12:347.

Lopez, M., Gonzalez-Barrera, A., and Patten, E. (2013). Closing the digital divide: Latinos and technology adoption. Technical report, Pew Research Center.

Lorca, A. and Sun, X. (2015). Adaptive robust optimization with dynamic uncertainty sets for multi-period economic dispatch under significant wind. *IEEE Transactions on Power Systems*, 30(4):1702–1713.

Mason, J., Denton, B., Smith, S., and Shah, N. (2013). Using electronic health records to monitor and improve adherence to medication. Working paper.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

McDonald, H., Garg, A., and Haynes, R. (2002). Interventions to enhance patient adherence to medication prescriptions: scientific review. *Journal of the American Medical Association*, 288(22):2868–2879.

Metz, C., Herman, B., and Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in Medicine*, 17(9):1033–1053.

Mifflin, M., St Jeor, S., Hill, L., Scott, B., Daugherty, S., and Koh, Y. (1990). A new predictive equation for resting energy expenditure in healthy individuals. *The American Journal of Clinical Nutrition*, 51(2):241–247.

Mignone, D., Ferrari-Trecate, G., and Morari, M. (2000). Stability and stabilization of piecewise affine and hybrid systems: An lmi approach. In *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, volume 1, pages 504–509. IEEE.

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y. (2017a). Behavioral analytics for myopic agents. *arXiv preprint arXiv:1702.05496.*

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y. (2017b). Non-stationary bandits with habituation and recovery dynamics. *arXiv preprint arXiv:1707.08423*.

Mohsenian-Rad, A. and Leon-Garcia, A. (2010). Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE transactions on Smart Grid*, 1(2):120–133.

Molderink, A., Bakker, V., Bosman, M., Hurink, J., and Smith, G. (2010). Management and control of domestic smart grid technology. *IEEE transactions on Smart Grid*, 1(2):109–119.

Moore, J. and Bard, J. (1992). An algorithm for the discrete bilevel programming problem. *Naval Research Logistics*, 39(3):419?435.

Murphy, S. A. and Vaart, A. W. V. D. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

Negoescu, D., Bimpikis, K., Brandeau, M., and Iancu, D. (2014). Dynamic learning of patient response types: An application to treating chronic diseases.

Nelles, O. (2001). *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Engineering online library. Springer.

O'Neill, D., Levorato, M., Goldsmith, A., and U. Mitra, U. (2010). Residential demand response using reinforcement learning. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 409–414. IEEE.

O'Reilly, G. and Spruijt-Metz, D. (2013). Current mHealth technologies for physical activity assessment and promotion. *American Journal of Preventive Medicine*, 45(4):501–507.

Osband, I., Blundell, C., Pritzel, A., and Roy, B. V. (2016). Deep exploration via bootstrapped DQN. *arXiv preprint arXiv:1602.04621*.

Osband, I. and Roy, B. V. (2015). Bootstrapped Thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*.

Ouattara, A. and Aswani, A. (2018). Duality approach to bilevel programs with a convex lower level. In *Proceedings of the American Control Conference*. To appear.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., and Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. *European Journal of Operational Research*, 266(2):639–651.

Pagoto, S., Schneider, K., Jojic, M., DeBiasse, M., and Mann, D. (2013). Evidence–based strategies in weight–loss mobile apps. *American Journal of Preventive Medicine*, 45(5):576–582.

Palensky, P. and Dietrich, D. (2011). Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE transactions on industrial informatics*, 7(3):381–388.

Papadimitriou, C. and Tsitsiklis, J. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305.

PG&E (2016). Save energy and money.

Portnoy, F. and Marchionini, G. (2010). Modeling the effect of habituation on banner blindness as a function of repetition and search type: Gap analysis for future work. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 4297–4302. ACM.

PwC, H. (2014). Health wearables: Early days. *Pricewaterhousecoopers, Top Health Industry Issues. Wearable Devices.*

Qu, X. and Keener, R. (2011). Theoretical statistics: Topics for a core course.

Radner, R. (1985). Repeated principal-agent games with discounting. *Econometrica*, 53(5):1173–1198.

Ralphs, T. and Hassanzadeh, A. (2014). On the value function of a mixed integer linear optimization problem and an algorithm for construction. Technical report.

Ratliff, L., Dong, R., Ohlsson, H., and Sastry, S. (2014). Incentive design and utility learning via energy disaggregation. *IFAC Proceedings Volumes*, 47(3):3158–3163.

Richter, W. (2017). Online ad revenues are surging, but 2 companies are getting most of the spoils.

Rockafellar, R. and Wets, R.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Rosenbaum, L. (2016). Should you really take 10,000 steps a day?

Rosenbaum, M., Hirsch, J., Gallagher, D. A., and Leibel, R. L. (2008). Long-term persistence of adaptive thermogenesis in subjects who have maintained a reduced body weight–. *The American journal of clinical nutrition*, 88(4):906–912.

Russo, D. and Roy, B. V. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Russo, D. and Roy, B. V. (2016). An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30.

Ryzhov, I. and Powell, W. (2011). Information collection on a graph. *Operations Research*, 59(1):188–201.

Ryzhov, I., Powell, W., and Frazier, P. (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195.

Saghezchi, F., Saghezchi, F., Nascimento, A., and Rodriguez, J. (2015). Game-theoretic based scheduling for demand-side management in 5g smart grids. In *2015 IEEE Symposium on Computers and Communication (ISCC)*, pages 8–12. IEEE.

Sahin, F. and Robinson, E. (2002). Flow coordination and information sharing in supply chains: review, implications, and directions for future research. *Decision sciences*, 33(4):505–536.

Samadi, P., Mohsenian-Rad, A., Schober, R., Wong, V., and Jatskevich, J. (2010). Optimal real-time pricing algorithm based on utility maximization for smart grid. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 415–420. IEEE.

Samadi, P., Mohsenian-Rad, H., Schober, R., and Wong, V. (2012). Advanced demand side management for the future smart grid using mechanism design. *IEEE Transactions on Smart Grid*, 3(3):1170–1180.

Sattelmair, J., Pertman, J., Ding, E., Kohl, H., Haskell, W., and Lee, I. (2011). Dose response between physical activity and risk of coronary heart disease. *Circulation*, pages CIRCULATIONAHA–110.

Savelsbergh, M. and Smilowitz, K. (2016). Stratified patient appointment scheduling for mobile community-based chronic disease management programs. *IIE Transactions on Healthcare Systems Engineering*, 6(2):65–78.

Schell, G., Marrero, W., Lavieri, M., Sussman, J., and Hayward, R. (2016). Data-driven markov decision process approximations for personalized hypertension treatment planning. *MDM Policy & Practice*, 1(1):2381468316674214.

Schoeller, D., Bandini, L., and Dietz, W. (1990). Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. *Canadian Journal of Physiology and Pharmacology*, 68(7):941–949.

Severini, T. (1999). On the relationship between bayesian and non-bayesian elimination of nuisance parameters. *Statistica Sinica*, 9:713–724.

Severini, T. and Wong, W. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics*, 20(4):1768–1802.

Shapiro, A. (1993). Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845.

Stackelberg, H. (1952). *The theory of market economy.* Oxford University Press.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Torres, F. (1991). Linearization of mixed-integer products. *Mathematical Programming*, 49(1–3):427–428.

Vahn, G. (2015). The data-driven (s, s) policy: Why you can have confidence in censored demand data. *Available at SSRN.*

Valero-Elizondo, J., Salami, J., Osondu, C., Latif, M., A, A., Spatz, E., Rana, J., Virani, S., Blankstein, R., Blaha, M., Veledar, E., and Nasir, K. (2016). Abstract 146: Drivers of healthcare costs among adults with obesity in united states: 2012 medical expenditure panel survey. *Circulation: Cardiovascular Quality and Outcomes*, 9(Suppl 2):A146.

Vielma, J. (2015). Mixed integer linear programming formulation techniques. *SIAM Review*, 57(1):3–57.

Vogel, S. and Lachout, P. (2003a). On continuous convergence and epi-convergence of random functions. part i: Theory and relations. *Kybernetika*, 39(1):[75]–98.

Vogel, S. and Lachout, P. (2003b). On continuous convergence and epi-convergence of random functions. part ii: Sufficient conditions and applications. *Kybernetika*, 39(1):99–118.

Wainwright, J. (2015). High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley.*

Wang, H., Zheng, B., Yoon, S. W., and Ko, H. S. (2017). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research.*

Wang, W.-Y. and Gupta, D. (2011). Adaptive appointment systems with patient preferences. *M&SOM*, 13:373.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298.

Wijaya, T., Papaioannou, T., Liu, X., and Aberer, K. (2013). Effective consumption scheduling for demand-side management in the smart grid using non-uniform participation rate. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2013*, pages 1–8. IEEE.

Wolsey, L. and Nemhauser, G. (1999). *Integer and Combinatorial Optimization.* Wiley-Interscience.

Wu, S., Ell, K., Gross-Schulman, S., Sklaroff, L., Katon, W., Nezu, A., Lee, P.-J., Vidyanti, I., Chou, C.-P., and Guterman, J. (2013). Technology-facilitated depression care management among predominantly Latino diabetes patients within a public safety net care system: Comparative effectiveness trial design. *Contemporary Clinical Trials.*

Xie, J. and Frazier, P. (2013). Sequential bayes-optimal policies for multiple comparisons with a known standard. *Operations Research*, 61(5):1174–1189.

Xie, W., Zhao, Y., Jiang, Z., and Chow, P. (2016). Optimizing product service system by franchise fee contracts under information asymmetry. *Annals of Operations Research*, 240(2):709–729.

Yi-jun, L., Li-gang, C., and Wen-guo, A. (2010). Exploitation vs. exploration: Choosing keywords for search-based advertising services. In *Management Science and Engineering (ICMSE), 2010 International Conference on*, pages 1–8. IEEE.

Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., and Aswani, A. (2018). Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR Mhealth Uhealth*, 6(1):e28.

Zhou, Z., Bambos, N., and Glynn, P. (2016). *Dynamics on Linear Influence Network Games Under Stochastic Environments*, pages 114–126. Springer International Publishing.

# Appendix A

## A.1 Proofs of Propositions in Chapter 2

*Proof of Proposition 2.1.* The result can be found using direct computation. This is because as $\alpha \downarrow 0$, $\alpha^0 \to 1$ while $\alpha^k \to 0$ for all $k > 0$. □

*Proof of Proposition 2.2.* This follows by first substituting the linear weight dynamics (2.1) and then noting that (i) the only stochasticity is in $z_t$, (ii) $z_t$ is zero mean, (iii) $z_t$ is unobservable at time $t$ and cannot be used to make a decision at time $t$, and (iv) the terms involving $z_t$ have an expectation of zero since the decisions are independent of $z_t$. □

*Proof of Proposition 2.3.* Because the constraints in $\mathbf{U_{no\ goals}}$ can be eliminated by rewriting the problem as $(u_t, f_t) = \arg\max_{u,f}\ -(a \cdot w_t + b \cdot u_t + c \cdot f_t + k)^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t$, the KKT conditions consist of only the stationarity conditions and are given by (2.4). A minor note is that $s_t = s_0$ here, because there are no dynamics on $s_t$ when goals are not provided as in $\mathbf{U_{no\ goals}}$. □

*Proof of Proposition 2.4.* Computing optimality conditions for $\mathbf{U_{goals}}$ requires reformulation as a quadratic program (QP) by using $p_t \cdot (u_t - g_t)^- = -\max\{-p_t \cdot (u_t - g_t), 0\}$. This QP reformulation has a differentiable, strictly concave objective and satisfies the linear independence constraint qualification (LICQ), and so the KKT conditions are necessary and sufficient for optimality. The KKT conditions can be rewritten after some manipulation as the first two lines of (2.5) combined with the following logical conditions on the Lagrange multipliers: $\lambda_t^2 = p_t$ if $u_t < g_t$, $0 \leq \lambda_t^2 \leq p_t$ if $u_t = g_t$, and $\lambda_t^2 = 0$ if $u_t > g_t$. Finally, let $M$ be a constant such that $M \geq p_t$. Using a big-M formulation (Vielma, 2015), we can express these logical conditions as in (2.5). □

*Proof of Proposition 2.5.* The first inequality states $x_{t+1}^1 \in \{0, 1\}$ (which indicates if $u_{t+1} \leq g_{t+1}$) can only decrease from $x_{t+1}^1$ if the goal decreases ($g_{t+1} < g_t$) or there is an office visit ($d_{t+1} = 1$). Similarly, the second and third inequalities state $x_{t+1}^2, x_{t+1}^3 \in$

$\{0, 1\}$ can only increase from $x_t^2, x_t^3$ if the goal decreases ($g_{t+1} < g_t$) or there is an office visit ($d_{t+1} = 1$). □

*Proof of Proposition 2.6.* Recall that (2.3) has the nonconvex quadratic term $\mu \cdot \mathbb{1}(u_t \geq g_t)$. Using the integer variables $x_t^1, x_t^2, x_t^3$ from the integer-reformulated KKT conditions (2.5), we can express this as the bilinear term $\mu \cdot (1 - x_t^1)$. This term has the special structure of a binary variable multiplied by a continuous scalar, and so a standard exact-linearization approach (Glover, 1975, Torres, 1991) can be used to reformulate the dynamics on $p_t$ as in (2.7). □

*Proof of Proposition 2.7.* First, note that the objective of $\mathbf{P_{pl}}$, after computing its negative logarithm, is proportional to:

$$\begin{aligned}
&\sigma_1^{-1/2}\sum_{i=1}^{n_w}|\tilde{w}_{t_i} - w_{t_i}|+\sigma_2^{-1/2}\sum_{i=1}^{n_u}|\tilde{u}_{\tau_i} - u_{\tau_i}|+\sigma_3^{-1/2}\sum_{t=1}^{n}|z_t| \\
&+ 2^{-1/2}\sum_{X\in\{\mu,q,s_0,\beta_0,\delta_0\}}\sum_{i=1}^{m_x}\log \pi_i^x \cdot \mathbb{1}(h_i^x \leq X \leq h_{i+1}^x) \\
&+ 2^{-1/2}\sum_{X\in\{\beta,\delta\}}\sum_{k=0}^{n_d-1}\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}\log \pi_{i,j}^x \cdot \mathbb{1}(h_i^x \leq X_{k+1} \leq h_{i+1}^x) \cdot \mathbb{1}(\phi_i^x \leq X_k \leq \phi_{i+1}^x),
\end{aligned} \tag{A.1}$$

where we have used the factorization of $\hat{\psi}(\Theta)$, the equation for one-dimensional histograms, the equation for the conditional histograms, and the equation for $\log \psi(\tilde{W}, \tilde{U}|W, U, F, \Theta, C)$ from $\mathbf{P_{mle-milp}}$. By defining $y_i^x \in \{0, 1\}$ for parameters $X \in \{\mu, q, s_0, \beta_0, \delta_0\}$ and $y_{i,j}^{x,k} \in \{0, 1\}$ for parameters $X \in \{\beta, \delta\}$, we can rewrite the objective function as in the hypothesis of the proposition. □

## A.2 Derivation of Weight Dynamics

The Mifflin St Jeor Equation[1] states that the basal metabolic rate (BMR) in units of calories/day is $10w_t + 6.25h - 5a + \sigma$, where $w_t$ is weight in kilograms, $h$ is height in centimeters, $a$ is age in years, $\sigma = +5$ for males, and $\sigma = -161$ for females. Additionally, 2000 steps is roughly equal to walking one mile and consumes about 100 calories, largely independent of the height, weight, age, and gender of an individual[2]. Based

[1] Mifflin, M., St Jeor, S., Hill, L., Scott, B., Daugherty, S., & Koh, Y. (1990). A new predictive equation for resting energy expenditure in healthy individuals. The American Journal of Clinical Nutrition, 51, 241247.

[2] Hill, J., Wyatt, H., Reed, G., & Peters, J. (2003). Obesity and the environment: Where do we go from here? Science, 299, 853855.

on the conversion factor that 3500 calories is equivalent to 1 pound, and 1 pound is equivalent to 0.45 kilograms: let $c = -0.45/3500$. Then the weight dynamics are

$$\begin{aligned} w_{t+1} &= w_t + c \cdot \tfrac{100}{2000} \cdot u_t + c \cdot f_t + c \cdot (10w_t + 6.25h - 5a + s) \\ &= (1 + 10c) \cdot w_t + c \cdot \tfrac{100}{2000} \cdot u_t + c \cdot f_t + c \cdot (6.25h - 5a + s) \\ &= aw_t + bu_t + cf_t + k \end{aligned}$$

where $s = -6.4286 \times 10^{-4}$ for males, $s = 2.0700$ for females, and

$$\begin{aligned} a &= 0.9987 \\ b &= -6.4287 \times 10^{-6} \\ c &= -1.2857 \times 10^{-4} \\ k &= -8.0357 \times 10^{-4}h + 6.4286 \times 10^{-4}a + s. \end{aligned}$$

## A.3 Complete MILP Formulation for MLE

$$
\begin{aligned}
\min\ & \sigma_1^{-1/2} \textstyle\sum_{i=1}^{n_w} \xi_{w,i} + \sigma_2^{-1/2} \textstyle\sum_{i=1}^{n_u} \xi_{u,i} + \sigma_3^{-1/2} \textstyle\sum_{t=1}^{n} \xi_{z,t} \\
\text{s.t.}\ & 2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q = 0, && \text{for } t = 1, \ldots, m-1 \\
& 2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_0 = 0, && \text{for } t = 1, \ldots, m-1 \\
& 2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q - \lambda_t^2 = 0, && \text{for } t = m, \ldots, n \\
& 2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_t = 0, && \text{for } t = m, \ldots, n \\
& g_t - \epsilon - (g_t - \epsilon) \cdot x_t^1 \le u_t \le M + (g_t - \epsilon - M) \cdot x_t^1, && \text{for } t = m, \ldots, n \\
& (g_t - \epsilon) \cdot x_t^2 \le u_t \le M + (g_t + \epsilon - M) \cdot x_t^2, && \text{for } t = m, \ldots, n \\
& (g_t + \epsilon) \cdot x_t^3 \le u_t \le g_t + \epsilon + (M - g_t - \epsilon) \cdot x_t^3, && \text{for } t = m, \ldots, n \\
& 0 \le \lambda_t^2 \le p_t, && \text{for } t = m, \ldots, n \\
& p_t - M \cdot (1 - x_t^1) \le \lambda_t^2 \le M \cdot (1 - x_t^3), && \text{for } t = m, \ldots, n \\
& x_t^1, x_t^2, x_t^3 \in \{0, 1\}, && \text{for } t = m, \ldots, n \\
& x_t^1 + x_t^2 + x_t^3 = 1, && \text{for } t = m, \ldots, n \\
& w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k + z_t, && \text{for } t = 1, \ldots, n-1 \\
& s_{t+1} = \gamma \cdot (s_t - s_0) + s_0 - \beta_{t+1} \cdot d_{t+1}, && \text{for } t = m, \ldots, n-1 \\
& p_{t+1} \ge \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1}, && \text{for } t = m, \ldots, n-1 \\
& p_{t+1} \le \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + M \cdot (1 - x_t^1), && \text{for } t = m, \ldots, n-1 \\
& p_{t+1} \ge \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu - M x_t^1, && \text{for } t = m, \ldots, n-1 \\
& p_{t+1} \le \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu, && \text{for } t = m, \ldots, n-1 \\
& x_{t+1}^1 \ge x_t^1 - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1 \\
& x_{t+1}^2 \le x_t^2 + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1 \\
& x_{t+1}^3 \le x_t^3 + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1 \\
& -\xi_{w,i} \le \tilde{w}_{t_i} - w_{t_i} \le \xi_{w,i}, && \text{for } i = 1, \ldots, n_w \\
& -\xi_{u,i} \le \tilde{u}_{\tau_i} - u_{\tau_i} \le \xi_{u,i}, && \text{for } i = 1, \ldots, n_u \\
& -\xi_{z,t} \le z_t \le \xi_{z,t}, && \text{for } t = 1, \ldots, n
\end{aligned}
$$

## A.4 Complete MILP Formulation for Bayesian Posterior

$$
\begin{aligned}
&\ell(w_{t_f} = \omega) = \\
&\qquad \min\ \sigma_1^{-1/2} \textstyle\sum_{i=1}^{n_w} \xi_{w,i} + \sigma_2^{-1/2} \sum_{i=1}^{n_u} \xi_{u,i} + \sigma_3^{-1/2} \sum_{t=1}^{n} \xi_{z,t} + \\
&\qquad\qquad 2^{-1/2} \textstyle\sum_{X \in \{\mu, q, s_0, \beta_0, \delta_0\}} \sum_{i=1}^{m_x} \log \pi_i^x \cdot y_i^x + \\
&\qquad\qquad\qquad 2^{-1/2} \textstyle\sum_{X \in \{\beta, \delta\}} \sum_{k=0}^{n_d - 1} \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \log \pi_{i,j}^x \cdot y_{i,j}^{x,k} \\
&\qquad \text{s.t. } 2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q = 0, && \text{for } t = 1, \ldots, m-1 \\
&\qquad\quad 2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_0 = 0, && \text{for } t = 1, \ldots, m-1 \\
&\qquad\quad 2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q - \lambda_t^2 = 0, && \text{for } t = m, \ldots, n \\
&\qquad\quad 2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_t = 0, && \text{for } t = m, \ldots, n \\
&\qquad\quad g_t - \epsilon - (g_t - \epsilon) \cdot x_t^1 \leq u_t \leq M + (g_t - \epsilon - M) \cdot x_t^1, && \text{for } t = m, \ldots, n \\
&\qquad\quad (g_t - \epsilon) \cdot x_t^2 \leq u_t \leq M + (g_t + \epsilon - M) \cdot x_t^2, && \text{for } t = m, \ldots, n \\
&\qquad\quad (g_t + \epsilon) \cdot x_t^3 \leq u_t \leq g_t + \epsilon + (M - g_t - \epsilon) \cdot x_t^3, && \text{for } t = m, \ldots, n \\
&\qquad\quad 0 \leq \lambda_t^2 \leq p_t, && \text{for } t = m, \ldots, n \\
&\qquad\quad p_t - M \cdot (1 - x_t^1) \leq \lambda_t^2 \leq M \cdot (1 - x_t^3), && \text{for } t = m, \ldots, n \\
&\qquad\quad x_t^1, x_t^2, x_t^3 \in \{0, 1\}, && \text{for } t = m, \ldots, n \\
&\qquad\quad x_t^1 + x_t^2 + x_t^3 = 1, && \text{for } t = m, \ldots, n \\
&\qquad\quad w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k + z_t, && \text{for } t = 1, \ldots, n-1 \\
&\qquad\quad s_{t+1} = \gamma \cdot (s_t - s_0) + s_0 - \beta_{t+1} \cdot d_{t+1}, && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad p_{t+1} \geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1}, && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad p_{t+1} \leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + M \cdot (1 - x_t^1), && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad p_{t+1} \geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu - M x_t^1, && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad p_{t+1} \leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu, && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad x_{t+1}^1 \geq x_t^1 - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad x_{t+1}^2 \leq x_t^2 + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1 \\
&\qquad\quad x_{t+1}^3 \leq x_t^3 + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0), && \text{for } t = m, \ldots, n-1
\end{aligned}
\tag{A.2}
$$

$$-\,\xi_{w,i} \leq \tilde{w}_{t_i} - w_{t_i} \leq \xi_{w,i}, \qquad \text{for } i = 1, \dots, n_w$$

$$-\,\xi_{u,i} \leq \tilde{u}_{\tau_i} - u_{\tau_i} \leq \xi_{u,i}, \qquad \text{for } i = 1, \dots, n_u$$

$$-\,\xi_{z,t} \leq z_t \leq \xi_{z,t}, \qquad \text{for } t = 1, \dots, n$$

$$\textstyle\sum_{i=1}^{m_x} h_i^x \cdot y_i^x \leq X \leq \sum_{i=1}^{m_x} h_{i+1}^x \cdot y_i^x, \qquad \text{for } X \in \{\mu, q, s_0, \beta_0, \delta_0\}$$

$$y_i^x \in \{0, 1\},\ \forall i = 1, \dots, m_x, \qquad \text{for } X \in \{\mu, q, s_0, \beta_0, \delta_0\}$$

$$\textstyle\sum_{i=1}^{m_x} y_i^x = 1, \qquad \text{for } X \in \{\mu, q, s_0, \beta_0, \delta_0\}$$

$$\textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} h_{i,j}^x \cdot y_{i,j}^{x,k} \leq X_{k+1} \leq \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} h_{i+1}^{x,k} \cdot y_{i,j}^x,$$
$$\text{for } X \in \{\beta, \delta\},\ k = 0, \dots, n_d - 1$$

$$\textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \phi_{i,j}^x \cdot y_{i,j}^{x,k} \leq X_k \leq \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \phi_{i+1}^x \cdot y_{i,j}^{x,k},$$
$$\text{for } X \in \{\beta, \delta\},\ k = 0, \dots, n_d - 1$$

$$y_{i,j}^{x,k} \in \{0, 1\},\ \forall i = 1, \dots, m_x,\ j = 1, \dots, \eta_x,$$
$$\text{for } X \in \{\beta, \delta\},\ k = 0, \dots, n_d - 1$$

$$\textstyle\sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} y_{i,j}^{x,k} = 1,\ \text{for } X \in \{\beta, \delta\},\ k = 0, \dots, n_d - 1$$

$$w_{t_f} = \omega$$

## A.5 Raw ROC Curve

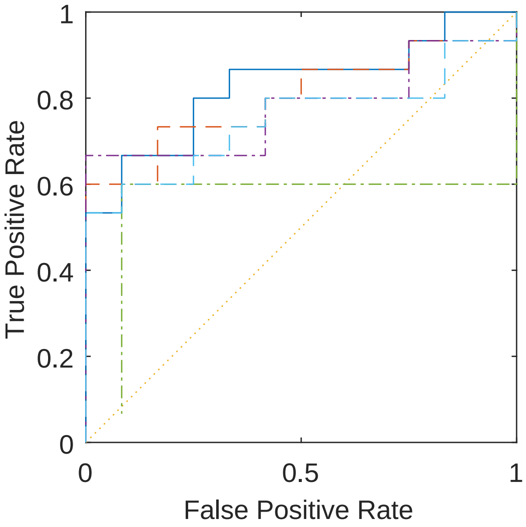


Figure A.1: Unsmoothed ROC computed using leave-one-out cross-validation for our predictive model with an empirical Bayesian prior (blue solid), our predictive model without a Bayesian prior (red dashed), linear SVM model (purple dash dot), decision tree model (green dashed dot), and logistic regression (cyan dashed) are compared.

# Appendix B

## B.1 Proofs of Propositions in Chapter 3

*Proof of Proposition 3.1.* The constraints $\theta_{t+1} = g(x_t, u_t, \theta_t, \pi_t)$ can be reformulated using Assumption 3.3 as

$$\begin{aligned} \theta_{t+1} &\leq G_i \cdot (x_t; u_t; \theta_t; \pi_t) + \xi_i + (1 - \iota_i) \cdot M \\ \theta_{t+1} &\geq G_i \cdot (x_t; u_t; \theta_t; \pi_t) + \xi_i - (1 - \iota_i) \cdot M \\ B_i \cdot (x_t; u_t; \theta_t; \pi_t) &\leq \psi_i + (1 - \iota_i) \cdot M \\ \iota_i &\in \{0, 1\} \end{aligned} \tag{B.1}$$

where $M > 0$ is a large-enough constant. Such a finite $M$ exists because $\mathcal{X}, \mathcal{U}, \Pi, \Theta$ are compact. Hence it suffices to show $u_t \in \text{argmax}\,\{f(x_{t+1}, u, \theta_t, \pi_t) \mid x_{t+1} = h(x_t, u),\ u \in \mathcal{U}\}$ can be represented (by its optimality condition) using a finite number of mixed integer linear constraints. Suppose $\mathcal{U} = \{u : \Xi u \leq \kappa\}$, where $\Xi$ is a matrix and $\kappa$ is a vector. Recall Assumption 3.2

$$f(x, u, \theta, \pi) = -(x; u)^T \cdot Q \cdot (x; u) + (\theta; \pi)^T \cdot H \cdot (x; u) + \sum_{i=1}^{K} \min_{j \in J_i} \{F_{i,j} \cdot (x; u; \theta; \pi) + \zeta_{i,j}\}. \tag{B.2}$$

We cannot characterize optimality by differentiating $f$ because it is generally not differentiable, but we can reformulate the maximization of (B.2) as the following convex quadratic program:

$$\begin{aligned} u_t \in \arg\max\ & -(x_{t+1}; u)^T \cdot Q \cdot (x_{t+1}; u) + (\theta_t; \pi_t)^T \cdot H \cdot (x_{t+1}; u) + \sum_{i=1}^{K} w_i \\ \text{s.t.}\ & w_i \leq F_{i,j} \cdot (x_{t+1}; u; \theta_t; \pi_t) + \zeta_{i,j} \text{ for all } i, j \\ & \Xi u \leq \kappa \end{aligned} \tag{B.3}$$

Using Assumption 3.3, we can rewrite the above as

$$
\begin{aligned}
u_t \in \arg\max \ & -u^T \cdot (B; \mathbb{I})^T \cdot Q \cdot (B; \mathbb{I}) \cdot u + ((\theta_t; \pi_t)^T H - 2(Ax_t + k; 0)^T Q) \cdot (B; \mathbb{I}) \cdot u + \sum_{i=1}^{K} w_i \\
\text{s.t. } & w_i \leq F_{i,j} \cdot (B; \mathbb{I}; 0; 0) \cdot u + F_{i,j} \cdot (Ax_t + k; 0; \theta_t; \pi_t) + \zeta_{i,j} \text{ for all } i, j \\
& \Xi u \leq \kappa
\end{aligned}
\tag{B.4}
$$

where we have eliminated the constant $(\theta_t; \pi_t)^T \cdot H \cdot (Ax_t + k; 0)$ since $x_t, \theta_t, \pi_t$ are known to the agent. The above optimization problem is convex with a with a strictly concave objective function by assumption, and all constraints are linear for fixed $\pi_t$. Hence the optimality conditions for (B.4) can be characterized using the KKT conditions (Dempe, 2002, Boyd and Vandenberghe, 2004). Let $\lambda_{i,j}$ and $\mu$ be the Lagrange Multipliers for the first and second set of constraints given in (B.4), and note the KKT conditions are

$$
\begin{aligned}
& 2(B; \mathbb{I})^T \cdot Q \cdot (B; \mathbb{I}) \cdot u + (B; \mathbb{I})^T \cdot (2Q \cdot (Ax_t + k; 0) - H^T \cdot (\theta_t; \pi_t)) + \Xi^T \cdot \mu = \\
& \qquad \sum_{i=1}^{K} \sum_{j \in J_i} \lambda_{i,j} \cdot (B; \mathbb{I}; 0; 0)^T \cdot F_{i,j}^T \\
& \sum_{j \in J_i} \lambda_{i,j} = 1 \text{ for } i = 1, \ldots, K \\
& \lambda_{i,j} \cdot (w_i - F_{i,j} \cdot (B; \mathbb{I}; 0; 0) \cdot u + F_{i,j} \cdot (Ax_t + k; 0; \theta_t; \pi_t) + \zeta_{i,j}) = 0 \text{ for all } i, j \\
& w_i \leq F_{i,j} \cdot (B; \mathbb{I}; 0; 0) \cdot u + F_{i,j} \cdot (Ax_t + k; 0; \theta_t; \pi_t) + \zeta_{i,j} \text{ for all } i, j \\
& \Xi u \leq \kappa \text{ and } \lambda_{i,j} \geq 0 \text{ for all } i, j
\end{aligned}
\tag{B.5}
$$

Note that the only nonlinear conditions are those which represent complimentary slackness. However, these conditions can be reformulated as integer linear conditions by posing them as disjunctive constraints (Wolsey and Nemhauser, 1999): For sufficiently large $M$ – which exists because of the compactness of $\mathcal{X}, \mathcal{U}, \Pi, \Theta$ – the complimentary slackness conditions are

$$
\begin{aligned}
& \lambda_{i,j} \leq M \iota_{i,j} \text{ for all } i, j \\
& F_{i,j} \cdot (B; \mathbb{I}; 0; 0) \cdot u + F_{i,j} \cdot (Ax_t + k; 0; \theta_t; \pi_t) + \zeta_{i,j} \leq w_i + M \cdot (1 - \iota_{i,j}) \text{ for all } i, j \\
& \iota_{i,j} \in \{0, 1\} \text{ for all } i, j
\end{aligned}
\tag{B.6}
$$

This shows that feasible region of (3.7) of can be represented using a finite number of mixed integer linear constraints. □

*Proof of Proposition 3.4.* Let $(x_0^*, \theta_0^*)$ be the agent's true initial conditions, and observe that

$$\log \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) = \log \hat{p}(x_0^*, \theta_0^* | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) + \\ \sum_{i=0}^{n_x} \log \frac{p_\nu(\tilde{x}_{t_i} - D\overline{x}_{t_i})}{p_\nu(\tilde{x}_{t_i} - Dx_{t_i})} + \sum_{j=0}^{n_u} \log \frac{p_\omega(\tilde{y}_{\tau_j} - C\overline{u}_{\tau_j})}{p_\omega(\tilde{y}_{\tau_j} - Cu_{\tau_j})} + \log \frac{p(x_0, \theta_0)}{p(x_0^*, \theta_0^*)}, \quad \text{(B.7)}$$

where $x_t, u_t$ are the states and decisions under initial conditions $(x_0^*, \theta_0^*)$, and $\overline{x}_t, \overline{u}_t$ are the states and decisions under initial conditions $(x_0, \theta_0)$. But $\log \frac{p(x_0,\theta_0)}{p(x_0^*,\theta_0^*)}$ is a constant by assumption, and $\log \hat{p}(x_0^*, \theta_0^* | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \leq 0$ since $\hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \in [0, 1]$ by construction. So using Assumption 3.6 gives $\max_{\mathcal{E}(\delta)} \log \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \to -\infty$ for any $\delta > 0$ almost surely. Equivalently, $\max_{\mathcal{E}(\delta)} \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \to 0$ for any $\delta > 0$ almost surely. Thus for any $\delta > 0$ we have that

$$\hat{p}(\mathcal{E}(\delta) | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) = \textstyle\int_{\mathcal{E}(\delta)} \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \times dx_0 \times d\theta_0 \leq \\ \text{volume}(\mathcal{X} \times \Theta) \cdot \max_{\mathcal{E}(\delta)} \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \to 0 \quad \text{(B.8)}$$

almost surely. This proves the result since (B.8) holds almost surely for any $\delta > 0$. □

*Proof of Corollary 3.5.* Consider the events:

$$\begin{aligned} E_1 =& \{(\hat{x}_{0,T}, \hat{\theta}_{0,T}) \notin \mathcal{B}(x_0^*, \theta_0^*, \delta)\} \\ E_2 =& \{\max_{\mathcal{E}(\delta)} \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}) \\ & \geq \max_{\mathcal{B}(x_0^*, \theta_0^*, \delta)} \hat{p}(x_0, \theta_0 | \{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T})\} \end{aligned} \quad \text{(B.9)}$$

where $\mathcal{E}(\delta)$ is defined as before for some $\delta > 0$. Then observe that $E_1 \subset E_2$, therefore $p_{(x_0^*,\theta_0^*)}(E_1) \leq p_{(x_0^*,\theta_0^*)}(E_2)$. By Proposition 3.4 as $T \to \infty$, $p_{(x_0^*,\theta_0^*)}(E_2) \to 0$ hence $p_{(x_0^*,\theta_0^*)}(E_1) \to 0$. Thus the result of the corollary follows. □

*Proof of Corollary 3.6.* For the first result, note Proposition 3.1 implies the feasible region of (3.15) can be expressed as mixed integer linear constraints with respect to $(x_t, u_t, \theta_t, \pi_t)$. Thus $\varphi(\overline{x}_0, \overline{\theta}_0, \{\overline{\pi}_i\}_{i=0}^{T+n})$ is the value function of a MILP in which $x_0, \theta_0, \overline{\pi}_t$ belong to an affine term. Standard results (Ralphs and Hassanzadeh, 2014) imply the value function is lower semicontinuous with respect to $x_0, \theta_0, \{\pi_i\}_{i=T+1}^{T+n}$.

To show the second result, note that the problem of $\min\{\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in \Pi^n\}$ is equivalent to (3.15) but with removal of the constraints $\pi_t = \overline{\pi}_t$ for $t = T+1, \ldots, T+n$. And so the result follows by Proposition 3.1 and by recalling the assumptions on $\Pi$ and $\ell$. □

*Proof of Proposition 3.7.* We begin by proving the first result. By definition we have that

$$\begin{aligned}&\mathbb{E}[\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n})|\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T}] = \\ &\qquad \textstyle\int_{\mathcal{X}\times\Theta} \varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n}) p(x_0, \theta_0|\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \pi) \times dx_0 \times d\theta_0.\end{aligned} \tag{B.10}$$

Also, Proposition 3.4 implies the posterior $p(x_0, \theta_0|\{\tilde{x}_{t_i}\}_{i=0}^{n_x}, \{\tilde{y}_{\tau_i}\}_{i=0}^{n_u}, \{\pi_i\}_{i=0}^{T})$ is consistent, and thus becomes degenerate at $(x_0^*, \theta_0^*)$ in the limit. Hence the Dominated Convergence Theorem gives

$$\text{(B.10)} \xrightarrow{p} \textstyle\int_{\mathcal{X}\times\Theta} \varphi(x_0, \theta_0, \pi) \times \delta(x_0 - x_0^*) \times \delta(\theta_0 - \theta_0^*) \times dx_0 \times d\theta_0 = \varphi(x_0^*, \theta_0^*, \pi), \tag{B.11}$$

where in the equation above $\delta(\cdot)$ is the Dirac delta function.

For the second result, recall that Corollary 3.5 implies $(\hat{x}_{0,T}, \hat{\theta}_{0,T}) \xrightarrow{p} (x_0^*, \theta_0^*)$. And Corollary 3.6 gives that $\varphi(x_0, \theta_0, \{\pi_i\}_{i=0}^{T+n})$ is lower semicontinuous in $x_0, \theta_0, \{\pi_i\}_{i=T+1}^{T+n}$. The result then follows by direct application of Proposition 2.1.ii of (Vogel and Lachout, 2003b). □

*Proof of Theorem 3.8.* The result follows by combining the second part of our Proposition 3.7 with Theorem 4.3 from (Vogel and Lachout, 2003a). □

*Proof of Proposition 3.9.* Step 2 of ABMA is a MAP estimate, which can be computed by solving a single MILP by Corollary 3.3. A similar argument used to prove Corollary 3.6 shows that Steps 4 and 5 can be computed by solving a single MILP. Step 8 can be seen to be an ILP by construction. The remaining steps of ABMA are assignment steps and do not require solving any optimization problems. □

*Proof of Proposition 3.10.* Using the assumptions on separability of the joint loss function $\Phi$ (Assumption 3.9) and decomposibility on the incentive set $\Omega$ (Assumption 3.10), we have that (3.19) can be reformulated as

$$\begin{aligned}\min_{y_v^a, \forall v, a \in V\times\mathcal{A}} \quad & \textstyle\sum_{a\in\mathcal{A}} \sum_{v\in V} \phi_v^a \cdot y_v^a \\ \text{s.t.} \quad & \min\{\varphi^a(x_{0,T}, \theta_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \mid \{\pi_i\}_{i=T+1}^{T+n} \in S_v\} \leq \phi_v^a \\ & x_0^a = \overline{x}_0^a, \theta_0^a = \overline{\theta}_0^a, \{\pi_t^a\}_{t=0}^{T+n} = \{\overline{\pi}_t^a\}_{t=0}^{T+n} \text{ for all } a \\ & \textstyle\sum_{a\in\mathcal{A}} \sum_{v\in V} v \cdot y_v^a \leq \beta \\ & \textstyle\sum_{v\in V} y_v^a = 1 \text{ for } a \in \mathcal{A} \\ & y_v^a \in \{0, 1\}\end{aligned} \tag{B.12}$$

Since $\phi_v^a \cdot y_v^a$ is the product of a continuous and binary decision variable, standard integer programming reformulation techniques allow us to reformulate the above as

$$\begin{aligned}
\min_{y_v^a, \forall v,a \in V \times \mathcal{A}} \quad & \textstyle\sum_{a\in\mathcal{A}} \sum_{v\in V} z_v^a \\
\text{s.t.} \quad & \varphi^a(x_{0,T}, \theta_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \leq \phi_v^a \\
& \{\pi_i\}_{i=T+1}^{T+n} \in S_v \\
& x_0^a = \overline{x}_0^a, \theta_0^a = \overline{\theta}_0^a, \{\pi_t^a\}_{t=0}^{T+n} = \{\overline{\pi}_t^a\}_{t=0}^{T+n} \text{ for all } a \\
& \textstyle\sum_{a\in\mathcal{A}} \sum_{v\in V} v \cdot y_v^a \leq \beta \\
& \textstyle\sum_{v\in V} y_v^a = 1 \text{ for } a \in \mathcal{A} \\
& y_v^a \in \{0,1\} \\
& z_v^a \geq \phi_v^a - M \cdot (1 - y_v^a) \\
& z_v^a \leq \phi_v^a + M \cdot (1 - y_v^a) \\
& z_v^a \leq M \cdot y_v^a \\
& z_v^a \geq -M \cdot y_v^a
\end{aligned} \tag{B.13}$$

where $M > 0$ is a large-enough constant. Such a finite $M$ exists because $\mathcal{X}, \mathcal{U}, \Pi, \Theta$ are compact, and because $\ell^a$ is representable by a finite number of mixed integer linear constraints. Since Corollary 3.6 (and its proof) implies we can represent $\varphi^a(x_{0,T}, \theta_{0,T}, \{\pi_i\}_{i=0}^{T+n}) \leq \phi_v^a$ and $\{\pi_i\}_{i=T+1}^{T+n} \in S_v$ by mixed integer linear constraints, this means we can reformulate (3.19) as a MILP with linear constraints that are affine in ($\overline{x}_0^a, \overline{\theta}_0^a, \{\overline{\pi}_i^a\}_{i=0}^{T+n}$ for $a \in \mathcal{A}$). And so standard results (Ralphs and Hassanzadeh, 2014) imply its value function is lower semicontinuous with respect to these variables, which is our first result. The second result follows by noting $\min\{\Phi(x_0^a, \theta_0^a, \{\pi_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n} \text{ for } a \in \mathcal{A}\} \in \Omega\}$ is equivalent to (3.19) but with removal of the constraints $\pi_t^a = \overline{\pi}_t^a$ for $t = T+1, \ldots, T+n$. □

*Proof of Theorem 3.12.* Corollary 3.5 implies $(\hat{x}_{0,T}^a, \hat{\theta}_{0,T}^a) \xrightarrow{p} (x_0^{*,a}, \theta_0^{*,a})$, and Corollary 3.6 states $\Phi$ is lower semicontinuous in its arguments. This means $\Phi(\hat{x}_0^a, \hat{\theta}_0^a, \{\pi_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A})$ is a lower semicontinuous approximation to $\Phi(x_0^{*,a}, \theta_0^{*,a}, \{\pi_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A})$ by Proposition 2.1.ii of (Vogel and Lachout, 2003b). But Corollary 3.11 shows that

$$\{\pi_{ABMA}^a(T) \text{ for } a \in \mathcal{A}\} \in \arg\min\{\Phi(\hat{x}_0^a, \hat{\theta}_0^a, \{\pi_i^a\}_{i=0}^{T+n} \text{ for } a \in \mathcal{A}) \mid \{\{\pi_t^a\}_{t=T+1}^{T+n} \text{ for } a \in \mathcal{A}\} \in \Omega\}. \tag{B.14}$$

This means that the result follows by applying Theorem 4.3 from (Vogel and Lachout, 2003a). □

## B.2 Complete MILP Formulation of MLE Problem

$$\min \frac{\sqrt{2}}{\sigma_1}\sum_{i=1}^{n_w}\zeta_{w,t_i} + \frac{\sqrt{2}}{\sigma_2}\sum_{i=1}^{n_s}\zeta_{s,t_i} \tag{B.15}$$

$$\text{s.t.} \quad -\zeta_{w,t_i} \leq \tilde{w}_{t_i} - w_{t_i} \leq \zeta_{w,t_i} \qquad \forall 1 \leq t_i \leq n_w \tag{B.16}$$

$$-\zeta_{s,t_i} \leq \tilde{s}_{\tau_i} - s_{\tau_i} \leq \zeta_{s,t_i} \qquad \forall 1 \leq t_i \leq n_s \tag{B.17}$$

$$-b(aw_1 + bs_0 + cf_0 + k) - r_s(s_0 - s_b) = 0 \tag{B.18}$$

$$-(aw_1 + bs_0 + cf_0 + k) - r_f(f_0 - F_{b,1}) = 0 \tag{B.19}$$

$$-b(aw_t + bs_t + cf_t + k) - r_s(s_t - s_b) = 0 \qquad \forall 1 \leq t \leq m \tag{B.20}$$

$$-(aw_t + bs_t + cf_t + k) - r_f(f_t - f_{b,t}) = 0 \qquad \forall 1 \leq t \leq m \tag{B.21}$$

$$-2b(aw_t + bs_t + cf_t + k) - 2r_s(s_t - s_b) + \lambda_{1,t} = 0 \qquad \forall m \leq t \leq n \tag{B.22}$$

$$-2(aw_t + bs_t + cf_t + k) - 2(f_t - f_{b,t}) = 0 \qquad \forall m \leq t \leq n \tag{B.23}$$

$$(g_t - \epsilon) - Mx_{1,t} \leq s_t \leq g_t - \epsilon + M(1 - x_{1,t}) \qquad \forall m \leq t \leq n \tag{B.24}$$

$$(g_t - \epsilon) - M(1 - x_{2,t}) \leq s_t \leq g_t + \epsilon + M(1 - x_{2,t}) \qquad \forall m \leq t \leq n \tag{B.25}$$

$$(g_t + \epsilon) - M(1 - x_{3,t}) \leq s_t \leq g_t + \epsilon + Mx_{3,t} \qquad \forall m \leq t \leq n \tag{B.26}$$

$$p_t - M(1 - x_{t,1}) \leq \lambda_{1,t} \leq M(1 - x_{3,t}) \qquad \forall m \leq t \leq n \tag{B.27}$$

$$0 \leq \lambda_{1,t} \leq p_t \qquad \forall m \leq t \leq n \tag{B.28}$$

$$\tag{B.29}$$

$$x_{t,1} + x_{t,2} + x_{t,3} = 1 \qquad \forall m \leq t \leq n \tag{B.30}$$

$$p_{t+1} \geq \gamma p_t + \delta d_{t+1} \qquad \forall m \leq t \leq n \tag{B.31}$$

$$p_{t+1} \leq \gamma p_t + \delta d_{t+1} + M(1 - x_{1,t}) \qquad \forall m \leq t \leq n \tag{B.32}$$

$$p_{t+1} \geq \gamma p_t + \delta d_{t+1} + \mu - Mx_{1,t} \qquad \forall m \leq t \leq n \tag{B.33}$$

$$p_{t+1} \leq \gamma p_t + \delta d_{t+1} + \mu \qquad \forall m \leq t \leq n \tag{B.34}$$

$$F_{b,t+1} = (1 - \alpha)F_{b,t} + \alpha f_{b,t} \qquad \forall t \tag{B.35}$$

$$f_{b,t+1} = \gamma(f_{b,t} - F_{b,t}) + F_{b,t} - \beta d_t \qquad \forall t \tag{B.36}$$

$$x_{t+1,1} \geq x_{t,1} - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.37}$$

$$x_{t+1,2} \leq x_{t,2} - d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.38}$$

$$x_{t+1,3} \leq x_{t,3} - d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.39}$$

$$w_t, s_t, f_t, F_{b,t}, f_{b,t}, p_t \geq 0 \qquad \forall t \tag{B.40}$$

$$x_{t,1}, x_{t,2}, x_{t,3} \in \mathbb{B} \qquad \forall t \tag{B.41}$$

## B.3 Complete MILP Formulation of MAP Problem

$$
\begin{align}
\min\ & \frac{\sqrt{2}}{\sigma_1}\sum_{i=1}^{n_w}\zeta_{w,t_i} + \frac{\sqrt{2}}{\sigma_2}\sum_{i=1}^{n_s}\zeta_{s,t_i} - \sum_{x\in\theta}\sum_{i=1}^{m_x} z_{i,x}\log\phi_i^x \tag{B.42}\\
\text{s.t.}\ & -\zeta_{w,t_i} \le \tilde{w}_{t_i} - w_{t_i} \le \zeta_{w,t_i} & \forall 1 \le t_i \le n_w \tag{B.43}\\
& -\zeta_{s,t_i} \le \tilde{s}_{\tau_i} - s_{\tau_i} \le \zeta_{s,t_i} & \forall 1 \le t_i \le n_s \tag{B.44}\\
& -b(aw_1 + bs_0 + cf_0 + k) - r_s(s_0 - s_b) = 0 \tag{B.45}\\
& -(aw_1 + bs_0 + cf_0 + k) - r_f(f_0 - F_{b,1}) = 0 \tag{B.46}\\
& -b(aw_t + bs_t + cf_t + k) - r_s(s_t - s_b) = 0 & \forall 1 \le t \le m \tag{B.47}\\
& -(aw_t + bs_t + cf_t + k) - r_f(f_t - f_{b,t}) = 0 & \forall 1 \le t \le m \tag{B.48}\\
& -2b(aw_t + bs_t + cf_t + k) - 2r_s(s_t - s_b) + \lambda_{1,t} = 0 & \forall m \le t \le n \tag{B.49}\\
& -2(aw_t + bs_t + cf_t + k) - 2(f_t - f_{b,t}) = 0 & \forall m \le t \le n \tag{B.50}\\
& (g_t - \epsilon) - Mx_{1,t} \le s_t \le g_t - \epsilon + M(1 - x_{1,t}) & \forall m \le t \le n \tag{B.51}\\
& (g_t - \epsilon) - M(1 - x_{2,t}) \le s_t \le g_t + \epsilon + M(1 - x_{2,t}) & \forall m \le t \le n \tag{B.52}\\
& (g_t + \epsilon) - M(1 - x_{3,t}) \le s_t \le g_t + \epsilon + Mx_{3,t} & \forall m \le t \le n \tag{B.53}\\
& p_t - M(1 - x_{t,1}) \le \lambda_{1,t} \le M(1 - x_{3,t}) & \forall m \le t \le n \tag{B.54}\\
& 0 \le \lambda_{1,t} \le p_t & \forall m \le t \le n \tag{B.55}\\
& x_{t,1} + x_{t,2} + x_{t,3} = 1 & \forall m \le t \le n \tag{B.56}\\
& p_{t+1} \ge \gamma p_t + \delta d_{t+1} & \forall m \le t \le n \tag{B.57}\\
& p_{t+1} \le \gamma p_t + \delta d_{t+1} + M(1 - x_{1,t}) & \forall m \le t \le n \tag{B.58}\\
& p_{t+1} \ge \gamma p_t + \delta d_{t+1} + \mu - Mx_{1,t} & \forall m \le t \le n \tag{B.59}\\
& p_{t+1} \le \gamma p_t + \delta d_{t+1} + \mu & \forall m \le t \le n \tag{B.60}\\
& \tag{B.61}
\end{align}
$$

$$F_{b,t+1} = (1-\alpha)F_{b,t} + \alpha f_{b,t} \qquad \forall t \tag{B.62}$$

$$f_{b,t+1} = \gamma(f_{b,t} - F_{b,t}) + F_{b,t} - \beta d_t \qquad \forall t \tag{B.63}$$

$$x_{t+1,1} \geq x_{t,1} - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.64}$$

$$x_{t+1,2} \leq x_{t,2} - d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.65}$$

$$x_{t+1,3} \leq x_{t,3} - d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \qquad \forall m \leq t \leq n \tag{B.66}$$

$$z_{i,x} h^x_{lb,i} \leq x_i \leq z_{i,x} h^x_{ub,i} \qquad \forall x \forall i \tag{B.67}$$

$$\sum_{i=1}^{m_x} z_{i,x} = 1 \qquad \forall x \forall i \tag{B.68}$$

$$\sum_{i=1}^{m_x} x_i = x \qquad \forall x \forall i \tag{B.69}$$

$$z_{i,x} \in \mathbb{B} \qquad \forall x \forall i \tag{B.70}$$

$$w_t, s_t, f_t, F_{b,t}, f_{b,t}, p_t \geq 0 \qquad \forall t \tag{B.71}$$

$$x_{t,1}, x_{t,2}, x_{t,3} \in \mathbb{B} \qquad \forall t \tag{B.72}$$

## B.4 Complete MILP Formulation of Personalized Treatment Plan Design Problem

$$\min \; w_n \tag{B.73}$$

$$\text{s.t. } F_{b,t+1} = (1-\alpha)F_{b,t} + \alpha f_{b,t} \qquad \forall t > T \tag{B.74}$$

$$f_{b,t+1} = \gamma(f_{b,t} - F_{b,t}) + F_{b,t} - \hat{\beta} d_t \qquad \forall t > T \tag{B.75}$$

$$p_{t+1} = \gamma p_t + \hat{\delta} d_t + \hat{\mu}(1 - x_{1,t}) \qquad \forall t > T \tag{B.76}$$

$$d_t \le \mathbb{1}\left[\mathrm{mod}(t,7) = 1\right] \qquad \forall t > T \tag{B.77}$$

$$d_t \le 1 - d_\tau \quad \forall \tau > T, \tau + 1 \le t \le \tau + 6 \qquad \forall t > T \tag{B.78}$$

$$-2b(aw_t + bs_t + cf_t + k) - 2r_s(s_t - s_b) + \lambda_{1,t} + \lambda_{4,t} = 0 \qquad \forall t > T \tag{B.79}$$

$$-2(aw_t + bs_t + cf_t + k) - 2(f_t - f_{b,t}) + \lambda_{3,t} = 0 \qquad \forall t > T \tag{B.80}$$

$$(g_t - \epsilon) - Mx_{1,t} \le s_t \le g_t - \epsilon + M(1 - x_{1,t}) \qquad \forall t > T \tag{B.81}$$

$$(g_t - \epsilon) - M(1 - x_{2,t}) \le s_t \le g_t + \epsilon + M(1 - x_{2,t}) \qquad \forall t > T \tag{B.82}$$

$$(g_t + \epsilon) - M(1 - x_{3,t}) \le s_t \le g_t + \epsilon + Mx_{3,t} \qquad \forall t > T \tag{B.83}$$

$$p_t - M(1 - x_{t,1}) \le \lambda_{1,t} \le M(1 - x_{3,t}) \qquad \forall t > T \tag{B.84}$$

$$0 \le f_t \le M(1 - x_{f,t}) \qquad \forall t > T \tag{B.85}$$

$$0 \le s_t \le M(1 - x_{s,t}) \qquad \forall t > T \tag{B.86}$$

$$0 \le \lambda_{3,t} \le Mx_{f,t} \qquad \forall t > T \tag{B.87}$$

$$0 \le \lambda_{4,t} \le Mx_{s,t} \qquad \forall t > T \tag{B.88}$$

$$0 \le \lambda_{1,t} \le p_t \qquad \forall t > T \tag{B.89}$$

$$x_{t,1} + x_{t,2} + x_{t,3} = 1 \qquad \forall t > T \tag{B.90}$$

$$g_{t+1} - g_t \le M(1 - g_{ind,t}) \qquad \forall t > T \tag{B.91}$$

$$g_{t+1} - g_t \ge -Mg_{ind,t} \qquad \forall t > T \tag{B.92}$$

$$x_{t+1,1} \ge x_{t,1} - d_{t+1} - g_{ind,t} \qquad \forall t > T \tag{B.93}$$

$$x_{t+1,2} \le x_{t,2} - d_{t+1} + g_{ind,t} \qquad \forall t > T \tag{B.94}$$

$$x_{t+1,3} \le x_{t,3} - d_{t+1} + g_{ind,t} \qquad \forall t > T \tag{B.95}$$

$$g_{ind,t} \in \mathbb{B} \tag{B.96}$$

$$x_{t,1}, x_{t,2}, x_{t,3}, x_{f,t}, x_{s,t} \in \mathbb{B} \tag{B.97}$$

$$d_t = \bar{d}_t; g_t = \bar{g}_t; w_t = \hat{w}_t; s_t = \hat{s}_t; f_t = \hat{f}_t; \theta_t = \hat{\theta}_t \qquad \forall t \le T \tag{B.98}$$

## B.5 Benchmarking Performance Tables

| Average Runtimes for Candidate Treatment Plan Calculation (in seconds) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Date of Calculation During the Program | | | | | |
| | | 15 | 30 | 60 | 90 | 120 | Average |
| Visit Budget | 54 | 21.715 | 9.3363 | 8.8984 | 9.5618 | 10.003 | 11.903 |
| | 63 | 21.715 | 7.7624 | 14.714 | 18.736 | 16.692 | 15.924 |
| | 72 | 21.715 | 8.8894 | 21.239 | 26.186 | 16.091 | 18.824 |
| | 81 | 21.715 | 11.408 | 43.147 | 28.112 | 23.284 | 25.533 |
| | 90 | 21.715 | 10.274 | 23.935 | 13.421 | 33.6 | 20.589 |
| | 99 | 21.715 | 9.3366 | 26.251 | 14.16 | 19.855 | 18.263 |
| | 108 | 21.715 | 9.4194 | 24.208 | 12.774 | 21.757 | 17.975 |
| | 117 | 21.715 | 9.9031 | 22.962 | 9.9226 | 31.865 | 19.273 |
| | 126 | 21.715 | 10.217 | 21.749 | 10.199 | 28.664 | 18.509 |
| | 135 | 21.715 | 9.973 | 16.307 | 23.129 | 18.893 | 18.003 |
| | 144 | 21.715 | 11.208 | 14.626 | 14.851 | 8.9203 | 14.264 |
| | 153 | 21.715 | 12.978 | 13.913 | 13.892 | 8.0501 | 14.11 |
| | 162 | 21.715 | 12.403 | 14.674 | 14.434 | 10.965 | 14.838 |
| | 171 | 21.715 | 13.305 | 12.102 | 11.585 | 23.78 | 16.497 |
| | 180 | 21.715 | 14.879 | 12.116 | 11.964 | 18.473 | 15.829 |
| | 189 | 21.715 | 11.731 | 11.835 | 12.221 | 21.715 | 15.843 |
| | Average | 21.715 | 10.814 | 18.917 | 15.322 | 19.538 | 17.261 |

Table B.1

| Average Runtimes for MAP Calculation (in seconds) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Date of Calculation During the Program | | | | | |
| | | 15 | 30 | 60 | 90 | 120 | Average |
| Visit Budget | 54 | 16.365 | 11.416 | 9.537 | 11.479 | 7.6851 | 11.296 |
| | 63 | 16.365 | 13.249 | 12.816 | 17.157 | 11.493 | 14.216 |
| | 72 | 16.365 | 11.913 | 14.15 | 13.846 | 12.884 | 13.832 |
| | 81 | 16.365 | 17.448 | 16.11 | 11.936 | 17.237 | 15.819 |
| | 90 | 16.365 | 12.251 | 9.9218 | 8.9877 | 16.463 | 12.798 |
| | 99 | 16.365 | 9.9473 | 10.59 | 9.3482 | 13.666 | 11.983 |
| | 108 | 16.365 | 10.412 | 10.199 | 9.377 | 14.746 | 12.22 |
| | 117 | 16.365 | 10.696 | 9.5879 | 8.9027 | 23.187 | 13.748 |
| | 126 | 16.365 | 10.36 | 10.737 | 9.6123 | 23.524 | 14.12 |
| | 135 | 16.365 | 9.7085 | 10.241 | 12.753 | 18.221 | 13.458 |
| | 144 | 16.365 | 9.3022 | 13.258 | 12.135 | 11.654 | 12.543 |
| | 153 | 16.365 | 8.5183 | 11.297 | 11.527 | 11.152 | 11.772 |
| | 162 | 16.365 | 9.4638 | 8.4174 | 11.279 | 14.812 | 12.068 |
| | 171 | 16.365 | 8.0878 | 11.092 | 9.8865 | 18.003 | 12.687 |
| | 180 | 16.365 | 8.8929 | 7.5454 | 6.5939 | 12.739 | 10.427 |
| | 189 | 16.365 | 7.679 | 7.6401 | 6.9321 | 16.365 | 10.996 |
| | Average | 16.365 | 10.584 | 10.821 | 10.735 | 15.239 | 12.749 |

Table B.2

| Average Runtimes for Knapsack Calculation (in seconds) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Date of Calculation During the Program | | | | | |
| | | 15 | 30 | 60 | 90 | 120 | Average |
| Visit Budget | 54 | 0.21791 | 0.19278 | 0.14949 | 0.1666 | 0.19679 | 0.18471 |
| | 63 | 0.19295 | 0.16084 | 0.19613 | 0.23577 | 0.22246 | 0.20163 |
| | 72 | 0.21302 | 0.2516 | 0.17474 | 0.18052 | 0.3125 | 0.22648 |
| | 81 | 0.25345 | 0.3542 | 0.18128 | 0.17449 | 0.25997 | 0.24468 |
| | 90 | 0.21698 | 0.17154 | 0.168 | 0.16942 | 0.21394 | 0.18798 |
| | 99 | 0.17072 | 0.17607 | 0.15058 | 0.15699 | 0.20012 | 0.1709 |
| | 108 | 0.17086 | 0.17178 | 0.17235 | 0.15172 | 0.19029 | 0.1714 |
| | 117 | 0.17958 | 0.17073 | 0.16465 | 0.16171 | 0.20232 | 0.1758 |
| | 126 | 0.18599 | 0.24335 | 0.17326 | 0.16726 | 0.20706 | 0.19538 |
| | 135 | 0.20403 | 0.16143 | 0.16626 | 0.16681 | 0.19607 | 0.17892 |
| | 144 | 0.17733 | 0.14174 | 0.16112 | 0.1618 | 0.18708 | 0.16581 |
| | 153 | 0.18841 | 0.15411 | 0.17753 | 0.17965 | 0.18822 | 0.17758 |
| | 162 | 0.17892 | 0.2167 | 0.18539 | 0.17419 | 0.21756 | 0.19455 |
| | 171 | 0.1771 | 0.29787 | 0.21357 | 0.2248 | 0.21789 | 0.22625 |
| | 180 | 0.20535 | 0.18508 | 0.22285 | 0.16565 | 0.19392 | 0.19457 |
| | 189 | 0.18554 | 0.2104 | 0.19875 | 0.17776 | 0.20053 | 0.1946 |
| | Average | 0.19488 | 0.20376 | 0.1785 | 0.17595 | 0.21292 | 0.1932 |

Table B.3

# Appendix C

## C.1 Proofs of Propositions in Chapter 4

*Proof of Proposition 4.2:* To obtain this formulation first we can augment the objective function of the log-likelihood problem by adding the constant term $\log \sum_{t \in \mathcal{T}_a} p(r_{a,t}|\theta_a^*, x_{a,t}^*)$ and multiplying by the positive constant $\frac{1}{n(\mathcal{T}_a)}$ which does not change the value of the optimal solution. Next we use functional compositions to contract the dynamics and obtain an objective function which is explicitly a function of $\theta_a, x_{0,a}$. □

*Proof of Lemma 4.1:* We can see that this is the case by noting that by Assumption 4.4 we have that each of the log-likelihood ratios are Lipschitz with constant $L_p$. Since Lipschitz continuity is preserved by addition and averaging we note that the average of all of these log-likelihood ratios is also $L_p$-Lipschitz. Next we use the property that functional compositions of Lipschitz functions are Lipschitz with a constant equal to the product of their respective constants and the Lipschitz continuity is preserved through point wise maxima (Rockafellar and Wets, 2009). Since the absolute value function is 1-Lipschitz and we are performing maximization we have that $\phi$ is indeed $L_p$-Lipschitz with respect to the input sequence. □

*Proof of Lemma 4.2:* To show the first result we use a similar argument to that of the proof of Lemma 4.1 by showing that the likelihood is Lipschitz and then using the preservation of Lipschitz continuity across functional compositions. First consider $h_a^t(x)$. Using the definition of $h_a$ from (4.1) we observe that $h_a$ is the composition of a linear function with a projection operator onto the set $\mathcal{X}$. Since projections are 1-Lipschitz (Rockafellar and Wets, 2009) and by Assumption 4.5 $\|A_a\|_{op}< 1$ we have that with respect to $x, x' \in \mathcal{X}$ $\|h_a^t(x) - h_a^t(x')\|_2< \|x - x'\|_2$. Hence $h_a^t(x)$ is locally 1-Lipschitz continuous with respect to $x, t$. Next, applying Assumption 4.3 shows that since the likelihood ratio is $L_f$-Lipschitz with respect to its two inputs we simply have a composition of Lipschitz functions and the result follows.

To show the second result note that $\ell$ depends on $t$ only through the composite dynamics mapping $h_a^t$. By definition $h^t(x) \in \mathcal{X}$ which is a bounded set, we have that for any $t, t' \in \{1, ..., T\}$ $\|h_a^t(x) - h_a^{t'}(x)\|_2 \leq \mathrm{diam}(\mathcal{X})$, thus using Assumption 4.3 we obtain the desired result. □

*Proof of Lemma 4.3.* To prove this result we first bound the expectation by a Rademacher average (Bartlett and Mendelson, 2002) and then apply Dudley's Integral bound (Wainwright, 2015). First let us consider the explicit form of $\mathbb{E}\varphi(\{r_{a,t}\}_{t=1}^{n(\mathcal{T}_a)})$. Using an identically distributed sequence of rewards $\{r'_{a,t}\}_{t=1}^{n(\mathcal{T}_a)}$ which is independent of the observed sequence we see that

$$\begin{aligned}
&\mathbb{E} \sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} |\frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \\
&\qquad\qquad - \frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \theta_a, x_{a,0})| \\
&= \mathbb{E} \sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \mathbb{E}[\sum_{t \in \mathcal{T}_a} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \\
&\qquad\qquad - \sum_{t \in \mathcal{T}_a} \log \frac{p(r'_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r'_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} |\{r_{a,t}\}_{t=1}^{n(\mathcal{T}_a)}] \Big| \\
&\leq \mathbb{E} \sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \Big( \sum_{t \in \mathcal{T}_a} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \\
&\qquad\qquad - \sum_{t \in \mathcal{T}_a} \log \frac{p(r'_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r'_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \Big) \Big|.
\end{aligned} \tag{C.1}$$

Here the inequality follows from Jensen's Inequality (Qu and Keener, 2011). Let $\{\epsilon_t\}_{t=1}^{n(\mathcal{T}_a)}$ be a sequence of i.i.d. Rademacher random variables, which are independent of the observations $r_{a,t}, r'_{a,t}$, then through a symmetrization argument its clear that

$$\mathbb{E}\varphi(\{r_{a,t}\}_{t=1}^{n(\mathcal{T}_a)}) \leq 2\mathbb{E} \sup_{\theta_a, x_{a,0} \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a, h_a^t(x_{a,0}, \theta_a, \pi_1^t))} \Big|. \tag{C.2}$$

Since $x_{a,0}^*, \theta_a^*$ are constants we can use simplify the above expression using the notation introduced in Lemma 4.2 to $2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t \ell(\theta_a, x_{0,a}, t) \Big|$. We can bound this expression as follows

$$\begin{aligned}
&2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t \ell(\theta_a, x_{0,a}, t) \Big|, \\
&= 2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t (\ell(\theta_a, x_{0,a}, t) - \ell(\theta_a, x_{0,a}, 0) + \ell(\theta_a, x_{0,a}, 0)) \Big|, \\
&\leq 2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t (\ell(\theta_a, x_{0,a}, t) - \ell(\theta_a, x_{0,a}, 0)) \Big| \\
&\quad + 2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t \ell(\theta_a, x_{0,a}, 0) \Big|.
\end{aligned} \tag{C.3}$$

For our analysis we can consider each of these terms separately and bound them using Dudley's Integral Bound (Wainwright, 2015) and Lemmas 4.2,C.1. Consider the first term, note that by Lemma 4.2 we have that $|\ell(\theta_a, x_{0,a}, t) - \ell(\theta_a, x_{0,a}, 0)| \leq L_f \operatorname{diam}(\mathcal{X})$ and is contained in an $\ell_2$ ball of this radius, hence by Lemma C.1

$$\begin{aligned}
&2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t (\ell(\theta_a, x_{0,a}, t) - \ell(\theta_a, x_{0,a}, 0)) \Big|, \\
&\leq 8 \int_0^{L_f \operatorname{diam}(\mathcal{X})} \sqrt{\frac{\log \mathcal{N}(L_f \operatorname{diam}(\mathcal{X}) B_2, \alpha, \|\|_2)}{n(\mathcal{T}_a)}} d\alpha \leq 8 L_f \operatorname{diam}(\mathcal{X}) \sqrt{\frac{\pi}{n(\mathcal{T}_a)}}.
\end{aligned} \tag{C.4}$$

The last inequality follows from using a volume bound on the covering number and using integration by parts. Next consider the second term in (C.3), we can bound this term using a direct application of Dudley's entropy integral as follows

$$\begin{aligned}
2\mathbb{E} \sup_{\theta_a, x_a \in \Theta \times \mathcal{X}} \Big| \frac{1}{n(\mathcal{T}_a)} \sum_{t \in \mathcal{T}_a} \epsilon_t \ell(\theta_a, x_{0,a}, 0) \Big| &\leq 16\sqrt{2} \int_0^{\infty} \sqrt{\frac{\log 2\mathcal{N}(\alpha, \ell(\Theta \times \mathcal{X}), \|\|_2)}{n(\mathcal{T}_a)}} d\alpha, \\
&\leq 16\sqrt{2} \int_0^{\infty} \sqrt{\frac{\log 2\mathcal{N}(\frac{\alpha}{L_f}, \Theta \times \mathcal{X}, \|\|_2)}{n(\mathcal{T}_a)}} d\alpha.
\end{aligned} \tag{C.5}$$

Let $v_\ell B_2$ be the $\ell_2$ ball on $\mathbb{R}^{d_x + d_\theta}$ with radius $v_\ell = \operatorname{diam}(\mathcal{X} \times \Theta)$, then

$$\text{(C.5)} \leq 16\sqrt{2} \int_0^{\infty} \sqrt{\frac{\log 2\mathcal{N}(\frac{\alpha}{L_f}, B_\ell, \|\|_2)}{n(\mathcal{T}_a)}} d\alpha \leq 16\sqrt{2} \int_0^{\infty} \sqrt{\frac{\log 2(\frac{3 v_\ell L_f}{\alpha})^{d_x + d_\theta}}{n(\mathcal{T}_a)}} d\alpha \tag{C.6}$$

Solving the integral shows that (C.6) $\leq 48\sqrt{2}(2)^{\frac{1}{d_x + d_\theta}} L_f v_\ell \sqrt{\frac{\pi(d_x + d_\theta)}{n(\mathcal{T}_a)}}$. Hence the result follows. □

*Proof of Theorem 4.3:* Lemma 4.1 guarantees that the mapping $\varphi$ is Lispschitz continuous with respect to the observed rewards with parameter $L_p$, furthermore we have by Assumption 4.4 that the reward distributions are sub-Gaussian with parameter $\sigma^2$. By applying Theorem 1 from Kontorovich (2014) we obtain for $\xi > 0$:

$$\mathbb{P}\Big(\varphi(\{r_t\}_{t=1}^{n(\mathcal{T}_a)}) - \mathbb{E}\varphi(\{r_t\}_{t=1}^{n(\mathcal{T}_a)}) > \xi\Big) \leq \exp(\frac{-\xi^2 n(\mathcal{T}_a)}{2L_p^2\sigma^2}). \tag{C.7}$$

Hence, using the upper bound obtained from Lemma 4.3, we can substitute the result into the above equation giving the desired result. □

*Proof of Theorem 4.1:* Using Theorem 4.3 we know that with probability at least $1 - \exp(\frac{-\xi^2 n(\mathcal{T}_a)}{2L_p^2\sigma^2})$ we have:

$$\begin{aligned}\frac{1}{n(\mathcal{T}_a)} D_{a,\pi_1^T}(\theta_a^*, x_{a,0}^* || \hat{\theta}_a, \hat{x}_{a,0}) - \frac{1}{n(\mathcal{T}_a)} \sum_{t \in n(\mathcal{T}_a)} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\hat{\theta}_a, h_a^t(\hat{x}_{a,0}, \hat{\theta}_a, \pi_1^t))}& \\ \leq \frac{c_f(d_x, d_\theta)}{\sqrt{n(\mathcal{T}_a)}} + \xi.&\end{aligned} \tag{C.8}$$

Also since $\hat{\theta}_a, \hat{x}_a$ are minimizers of the empirical trajectory divergence implies that

$$\begin{aligned}\frac{1}{n(\mathcal{T}_a)} \sum_{t \in n(\mathcal{T}_a)} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\hat{\theta}_a, h_a^t(\hat{x}_{a,0}, \hat{\theta}_a, \pi_1^t))}& \\ \leq \frac{1}{n(\mathcal{T}_a)} \sum_{t \in n(\mathcal{T}_a)} \log \frac{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))}{p(r_{a,t}|\theta_a^*, h_a^t(x_{a,0}^*, \theta_a^*, \pi_1^t))} = 0.&\end{aligned} \tag{C.9}$$

Hence the desired result follows. □

*Proof of Proposition 4.6:* Recall that by definition $\mathbb{E}R_\Pi(T) = \sum_{t=1}^T g(\theta_{pi_t^*}, x_{pi_t^*}) - g(\theta_{pi_t}, x_{pi_t})$. Since by Assumption 4.3 we have that $g$ is $L_g$-Lipschitz then we have $\forall t$ that $g(\theta_{pi_t^*}, x_{pi_t^*}) - g(\theta_{pi_t}, x_{pi_t}) \leq L_g \|(\theta_{pi_t^*}, x_{pi_t^*}) - (\theta_{pi_t}, x_{pi_t})\| \leq L_g \operatorname{diam}(\mathcal{X} \times \Theta)\mathbb{P}(\pi_t \neq$

$\pi_t^*$). Hence

$$\begin{aligned}\mathbb{E}R_\Pi(T) &\leq L_g \operatorname{diam}(\mathcal{X}\times\Theta)\sum_{t=0}^{T}\mathbb{P}(\pi_t \neq \pi_t^*)\\ &= L_g \operatorname{diam}(\mathcal{X}\times\Theta)\sum_{t=0}^{T}\sum_{a\in\mathcal{A}}\mathbb{P}(\pi_t = a, a \neq \pi_t^*)\\ &= L_g \operatorname{diam}(\mathcal{X}\times\Theta)\sum_{a\in\mathcal{A}}\sum_{t=0}^{T}\mathbb{P}(\pi_t = a, a \neq \pi_t^*)\\ &= L_g \operatorname{diam}(\mathcal{X}\times\Theta)\sum_{a\in\mathcal{A}}\mathbb{E}\tilde{T}_a\end{aligned} \tag{C.10}$$

□

*Proof of Proposition 4.7:* We proceed to prove this proposition in a similar method to that presented in Auer et al. (2002b). Suppose that at time $t$, the ROGUE-UCB policy chooses $a \neq \pi_t^*$. If the upper confidence bounds hold then we observe that $g_{a,t}^{UCB} \geq g_{\pi_t^*,t}^{UCB} \geq g_{\pi_t^*,t}$. Also define the mapping $\psi_a(\gamma) = \max\{|g(\theta, h_a^t(x_0)) - g(\hat{\theta}_a, h_a^t(\hat{x}_{a,0}))|\colon \frac{1}{n(\mathcal{T}_a)}D_{a,\pi_1^T}(\theta, x_0||\hat{\theta}_a, \hat{x}_{a,0}) \leq \gamma\}$. Then clearly $g_{a,t}^{U}CB - g(\hat{\theta}_a, h_a^t(\hat{x}_{a,0})) \leq \psi_a(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}})$ and $g(\hat{\theta}_a, h_a^t(\hat{x}_{a,0})) - g_{a,t} \leq \psi_a(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}})$. Hence we have that $g_{a,t}^{UCB} \leq 2\psi(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}) + g_{a,t}$. Therefore $\psi(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}) \geq \frac{1}{2}(g_{\pi_t^*,t} - g_{a,t})$. By definition of $\epsilon_a$ we thus have that $\psi(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}) \geq \frac{\epsilon_a}{2}$. Therefore, by definition of $\delta_a$ we observe that $A(t)\sqrt{\frac{4\log t}{n(\mathcal{T}_a)}} \geq \delta_a$ and hence $n(\mathcal{T}_a) \leq \frac{4A(t)^2\log t}{\delta_a^2}$.

Now, consider $\tilde{T}_a$:

$$\tilde{T}_a = \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*\} \tag{C.11}$$

$$\begin{aligned} =&\sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) \leq \frac{4A(t)^2 \log t}{\delta_a^2}\} \\ &\qquad + \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \end{aligned} \tag{C.12}$$

$$\begin{aligned} \leq&\sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) \leq \frac{4A(|\mathcal{A}|)^2 \log T}{\delta_a^2}\} \\ &\qquad + \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \end{aligned} \tag{C.13}$$

$$\leq \frac{4\log(T)}{\delta_a^2} A(|\mathcal{A}|)^2 + \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \tag{C.14}$$

Observe that if we play sub optimal action $a$ at time $t$ this means we either severely over estimate the value of $g_{a,t}$, severely under estimate the value of $g_{\pi_t^*,t}$, or the two values are very close to each other. Hence

$$\begin{aligned} &\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \\ &\subseteq \underbrace{\{g_{a,t}^{UCB} - g_{a,t} > 2\psi_a(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}), n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\}}_{(a)} \\ &\cup \underbrace{\{g_{\pi_t^*,t} > g_{UCB}^{\pi_t^*,t}, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\}}_{(b)} \\ &\cup \underbrace{\{g_{\pi_t^*,t} - g_{a,t} \leq 2\psi_a(A(t)\sqrt{\frac{4\log(t)}{n(\mathcal{T}_a)}}), n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\}}_{(c)}. \end{aligned} \tag{C.15}$$

However, as we established in the beginning of the proof the event $(c) = \emptyset$. Also note that for events $(a), (b)$ to occur this would imply that $\theta_a, x_{a,0}$ and $\theta_{\pi_t^*}, x_{\pi_t^*,0}$ are not feasible points of their respective UCB deriving problems, hence

$$\begin{aligned}
&\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \\
&\subseteq \{\exists s < t: \ \frac{1}{s} D_{\pi_t^*, \pi_1^s}(\hat{\theta}_{\pi_t^*}, \hat{x}_{\pi_t^*,0} || \theta_{\pi_t^*}, x_{\pi_t^*,0}) > A(t)\sqrt{\frac{4\log(t)}{s}}\} \\
&\cup \{\exists s' < t: \ \frac{1}{s'} D_{a, \pi_1^{s'}}(\hat{\theta}_a, \hat{x}_{a,0} || \theta_a, x_{a,0}) > A(t)\sqrt{\frac{4\log(t)}{s'}}\} \\
&\subseteq \bigcup_{1 \leq s < t} \{\frac{1}{s} D_{\pi_t^*, \pi_1^s}(\hat{\theta}_{\pi_t^*}, \hat{x}_{\pi_t^*,0} || \theta_{\pi_t^*}, x_{\pi_t^*,0}) > A(t)\sqrt{\frac{4\log(t)}{s}}\} \\
&\qquad \bigcup_{1 \leq s' < t} \{\frac{1}{s'} D_{a, \pi_1^{s'}}(\hat{\theta}_a, \hat{x}_{a,0} || \theta_a, x_{a,0}) > A(t)\sqrt{\frac{4\log(t)}{s'}}\}.
\end{aligned} \tag{C.16}$$

Taking the expected value of $\tilde{T}_a$ we obtain

$$\begin{aligned}
&\mathbb{E}\tilde{T}_a \leq \frac{4\log(T)}{\delta_a^2} A(|\mathcal{A}|)^2 + \mathbb{E}\sum_{t=1}^{T} \mathbf{1}\{\pi_t = a, a \neq \pi_t^*, n(\mathcal{T}_a) > \frac{4A(t)^2 \log t}{\delta_a^2}\} \\
&\leq \frac{4\log(T)}{\delta_a^2} A(|\mathcal{A}|)^2 \\
&+ \sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{s'=1}^{t} \mathbb{P}(\frac{1}{s} D_{\pi_t^*, \pi_1^s}(\hat{\theta}_{\pi_t^*}, \hat{x}_{\pi_t^*,0} || \theta_{\pi_t^*}, x_{\pi_t^*,0}) > A(t)\sqrt{\frac{4\log(t)}{s}}) \\
&+ \sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{s'=1}^{t} \mathbb{P}(\frac{1}{s'} D_{a, \pi_1^{s'}}(\hat{\theta}_a, \hat{x}_{a,0} || \theta_a, x_{a,0}) > A(t)\sqrt{\frac{4\log(t)}{s'}}) \\
&\leq \frac{4\log(T)}{\delta_a^2} A(|\mathcal{A}|)^2 + 2\sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{s'=1}^{t} t^{-4} \leq \frac{4\log(T)}{\delta_a^2} A(|\mathcal{A}|)^2 + \frac{\pi^2}{3}.
\end{aligned} \tag{C.17}$$

Here the third inequality is derived by Theorem 4.1 and the final inequality by utilizing the solution to the Basel Problem (Rockafellar and Wets, 2009). Hence we obtain the desired result. □

*Proof of Theorem 4.5:* Using Proposition 4.6 we bound the expected regret as $\mathbb{E}R_\Pi(T) \leq L_g \operatorname{diam}(\mathcal{X} \times \Theta) \sum_{a \in \mathcal{A}} \mathbb{E}\tilde{T}_a$. Then applying the result of Proposition 4.7 we obtain the desired result. □

## C.2 Technical Metric Entropy Lemma

**Lemma C.1.** *Let $a \in A \subseteq \mathbb{R}^n$ such that $A$ is bounded and $K = \max_{a\in A} \frac{d(a,0)}{n}$ with respect to some metric $d$ and $\forall a \in A, \|a\|_2 \leq d(a,0)$ . Then for i.i.d Rademacher process $\{\epsilon_i\}_{i=1}^n$ :*

$$\mathbb{E} \sup_{a\in A} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right| \leq 4\int_0^K \sqrt{\frac{\log 2\mathcal{N}(\alpha, A, d)}{n}} d\alpha \tag{C.18}$$

*Proof:* We proceed to prove this result in a similar technique to that used by (Wainwright, 2015). Let $\bar{A} = A \cup A^-$ and $\{\hat{A}_i\}_{i=0}^N$ be a sequence of successively finer covers of set $\bar{A}$, such that $\hat{A}_i$ is an $\alpha_i$ cover of set $\bar{A}$ with respect to metric $d$ and $\alpha_i = 2^{-i}K$. Next, define a sequence of approximating vectors of $a$ and denote these by $\hat{a}_i$ such that for any two successive approximations $\hat{a}_i \in \hat{A}_i$ and $\hat{a}_{i-1} \in \hat{A}_{i-1}$ we have that $d(\hat{a}_i, \hat{a}_{i-1}) \leq \alpha_i$. Then observe we can rewrite $a$ as follows:

$$a = a + \hat{a}_N - \hat{a}_N = \hat{a}_0 + \sum_{i=1}^N (\hat{a}_i - \hat{a}_{i-1}) + a - \hat{a}_n \tag{C.19}$$

Observe that we can set $\hat{a}_0$ to the 0 vector since clearly a metric ball of radius $K$ will form a $K$ cover of set $A$. Hence we obtain:

$$\mathbb{E} \sup_{a\in A} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right| = \mathbb{E} \sup_{a\in A} \left|\frac{1}{n}\langle \epsilon, a\rangle\right| = \mathbb{E} \sup_{a\in \bar{A}} \frac{1}{n}\langle \epsilon, a\rangle = \mathbb{E} \sup_{a\in \bar{A}} \frac{1}{n}\langle \epsilon, \sum_{j=1}^N (\hat{a}_i - \hat{a}_{i-1}) + a - \hat{a}_N\rangle \tag{C.20}$$

$$\leq \mathbb{E}\sum_{j=1}^N \sup_{\hat{a}_j\in \hat{A}_j, \hat{a}_{j-1}\in \hat{A}_{j-1}} \langle \epsilon, \hat{a}_j - \hat{a}_{j-1}\rangle + \mathbb{E}\sup_{a\in \bar{A}} \langle \epsilon, a - \hat{a}_N\rangle \tag{C.21}$$

$$\leq \sum_{j=1}^N \alpha_i \sqrt{\frac{2\log|\hat{A}_j||\hat{A}_{j-1}|}{n}} + \alpha_N \tag{C.22}$$

Here the final inequality is obtained by applying the finite class lemma (Wainwright, 2015). Observe that $|\hat{A}_{j-1}| \leq |\hat{A}_{j-1}| = \mathcal{N}(\alpha_i, \bar{A}, d)$ and that by construction $\alpha_j = 2(\alpha_j - \alpha_{j+1})$. Hence:

$$\mathbb{E} \sup_{a\in A} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right| \leq \sum_{j=1}^N 4(\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log \mathcal{N}(\alpha_i, \bar{A}, d)}{n}} + \alpha_N \tag{C.23}$$

$$\leq 4\int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\frac{\log \mathcal{N}(\alpha, \bar{A}, d)}{n}} d\alpha + \alpha_N \to 4\int_0^K \sqrt{\frac{\log \mathcal{N}(\alpha, \bar{A}, d)}{n}} d\alpha \tag{C.24}$$

Note that $\mathcal{N}(\alpha, \bar{A}, d) \leq 2\mathcal{N}(\alpha, A, d)$ thus completing the proof. □