An Online Actor Critic Algorithm and a Statistical Decision Procedure for Personalizing Intervention

by

Huitian Lei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2016

Doctoral Committee:

Professor Susan A. Murphy, co-Chair
Assistant Professor Ambuj Tewari, co-Chair
Associate Professor Lu Wang
Assistant Professor Shuheng Zhou

2016

Dedication

To my mother

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

An Online Actor Critic Algorithm and a Statistical Decision Procedure for Personalizing Intervention

by

Huitian Lei

Chair: Professor Susan A. Murphy

Assistant Professor Ambuj Tewari

Increasing technological sophistication and widespread use of smartphones and wearable devices provide opportunities for innovative health interventions. An Adaptive Intervention (AI) personalizes the type, mode and dose of intervention based on users' ongoing performances and changing needs. A Just-In-Time Adaptive Intervention (JITAI) employs the real-time data collection and communication capabilities that modern mobile devices provide to adapt and deliver interventions in real-time. The lack of methodological guidance in constructing data-based high quality JITAI remains a hurdle in advancing JITAI research despite its increasing popularity. In the first part of the dissertation, we make a first attempt to bridge this methodological gap by formulating the task of tailoring interventions in real-time as a contextual bandit problem. Under the linear reward assumption, we choose the reward function (the "critic") parameterization separately from a lower dimensional parameterization of stochastic JITAIs (the "actor"). We provide an online actor critic algorithm that guides the construction and refinement of a JITAI. Asymptotic properties of the actor critic algorithm, including consistency, asymptotic distribution and regret bound of the optimal JITAI parameters are developed and tested by numerical experiments. We also present numerical experiment to test performance of the algorithm when assumptions

in the contextual bandits are broken. In the second part of the dissertation, we propose a statistical decision procedure that identifies whether a patient characteristic is useful for AI. We define a discrete-valued characteristic as useful in adaptive intervention if for some values of the characteristic, there is sufficient evidence to recommend a particular intervention, while for other values of the characteristic, either there is sufficient evidence to recommend a different intervention, or there is insufficient evidence to recommend a particular intervention.

CHAPTER 1

Introduction

Advanced technology in smartphones and mobile devices provide a great platform to deliver Just-In-Time Adaptive Interventions (JITAI). Adaptive intervention tailors the type, dosage or modality of the intervention according to patients' characteristics. JITAI is a real-time version of AI. [58] provides a definition for JITAI: *"JITAIs are interventions that are delivered in real-time, and are adapted to address the immediate and changing needs of individuals as they go about their daily lives."* Based on real-time information collected on the mobile devices, JITAI personalizes, in real-time, the type, mode and dose of intervention based on users' ongoing performances and changing needs and delivers the intervention on mobile devices. The real-time adaptation and delivery makes JITAI particular promising in facilitating behavioral change. Indeed JITAIs have received increasing popularity and have been used to support health behavior change in a variety of domain including physical activity [37, 19], eating disorder [4], drug abuse [72], alcohol use [89, 76, 31] smoking cessation [68], obesity/weight management [60] and other chronic disorders.

Despite the growing popularity of JITAI, there is a lack of guidance on constructing high-quality evidence-based JITAI. In fact, most of the JITAIs used in existing clinical trials are solely based on domain expertise. The *major* contribution of this dissertation is that we make a first step to bridge the gap between the enthusiasm to deliver intervention on mobile devices and the lack of statistical tools to guide the building to high-quality JITAI. To achieve our goal, we first propose a general framework for constructing high quality JITAI. We model the decision making problem, choosing (the dosage/type of) an intervention based on information collected on the mobile device, as a contextual bandit problem [43, 50]. Contextual bandit problem is a special type of sequential decision making problem where the decision maker (i) chooses an action at each round based on the *context* or side information and (ii) receives an reward/feedback that reflects the quality of the action under the context. Bandit problems have been widely applied in clinical trials, economics and portfolio designs and have recently found applications in mobile health [65]. We provide a brief review on multi-armed bandits and contextual bandits in section 1.2. We define

the optimal JITAI as the JITAI that maximizes the average reward subject to a stochasticity constraint. We propose an online actor critic algorithm for learning the optimal JITAI. Compared to offline learning, in online learning the data comes in in a sequential fashion and the estimated optimal JITAI gets updated at each decision point and will be used to choose an intervention at the next decision point. In the actor critic algorithm, the critic estimates the reward model; actor then updates its estimate to the optimal JITAI based on the estimated reward model. We derive asymptotic theory on the consistency and asymptotic normality of the estimated optimal JITAI. Asymptotic distribution of the estimated optimal JITAI can be used to construct statistical hypothesis test on whether a component of context is useful for tailoring intervention.

Often, the i.i.d. assumption in contextual bandits is fragile. Contexts may not be i.i.d. but are instead influenced by the context or the intervention at previous decision points. We conduct simulation studies to test the performance of the contextual bandit actor critic algorithm under a variety of simulation settings. Results from the experiments where contexts are i.i.d. are consistent with the asymptotic theory: bias in estimating optimal JITAI decreases to 0 as sample size increases. Results from the experiments where contexts follow a first degree auto-regressive process show that the bandit actor critic algorithm is robust to the dependency between contexts at different decision points. We also create simulation settings where the context is influenced by previous actions–in one setting through a learning effect and in the other setting through a burden effect. In both settings, we observe robustness of the algorithm when the effect of previous actions are small.

A *minor* contribution of the dissertation is that we introduce a statistical decision procedure for personalizing intervention. The decision procedure is used to decide whether a binary-valued patient characteristic is useful for personalizing decision making. We define a characteristic to be useful if at one level of the characteristic there is sufficient evidence to recommend a particular intervention while at the other level either there is sufficient evidence to recommend another intervention (qualitative interaction) or there is insufficient evidence to recommend a particular intervention (generalized qualitative interaction). The new definition is a generalization of the qualitative interaction [28] and recognizes the increased utility when patients are provided with freedom to choose an intervention. We propose a two stage multiple decision procedure that decides whether the evidence suggests a qualitative interaction, and if not, whether there is a generalized qualitative interaction.

This dissertation is organized as follows. In Chapter 2, we introduce the formation of the problem as a contextual bandit problem. Because of the nature of the our target application, we study a parametrized class of policies unlike most contextual bandit algorithms, which either maintain a finite class of policies or do not maintain a class of policies. By

adding a stochasticity constraint, our definition of optimality is different from the one used in existing literature. We present an actor critic contextual bandit algorithm for linear expected reward. We derive asymptotic theory on the consistency and asymptotic normality of the optimal JITAI. In Chapter 3, we present a comprehensive simulation study to investigate the performance of the actor critic algorithm under various generative models. In Chapter 4, we propose a multiple decision procedure for personalizing intervention.

1.1 A Review on Adaptive Intervention and Just-in-time Adaptive Intervention

Adaptive interventions are interventions in which the type or the dosage of the intervention offered to patients is individualized on the basis of patients characteristics or clinical presentation and can be repeatedly adjusted over time in response to their ongoing performance (see, for example, [10, 54]). This approach is based on the notion that patients differ in their responses to interventions: In order for an intervention to be most effective, it should be individualized and repeatedly adapted over time to individual progress. An adaptive intervention is a multi-stage process that can be operationalized via a sequence of decision rules that recommend when and how the intervention should be modified in order to maximize long-term primary outcomes. These recommendations are based not only on patients' characteristics but also on intermediate outcomes collected during the intervention, such as the patient?s response and adherence. Adaptive interventions are also known as dynamic treatment regimes [57, 70], adaptive treatment strategies [44, 56], multi-stage treatment strategies [80, 81] and treatment policies [52, 86, 87].

An adaptive intervention consists of four key elements. The first element is a sequence of critical decisions in a patient care. Critical decisions might concern which intervention to provide first and, if the initial intervention is unsuccessful, which intervention to provide second. In many settings, the risk of relapse or exacerbations is high; thus, critical decisions must be made even after an acute response has been achieved. These decisions may concern which maintenance intervention should be offered and whether and how signs of impending relapse should be monitored [55]. The second element is a set of possible intervention options at each critical decision point. Possible intervention options include different types of behavioral and pharmacological interventions, different modes of delivery, different combinations of interventions, different approaches to enhance engagement and adherence to the intervention, and different intervention timelines. The third element is a set of tailoring variables that is used to pinpoint when the intervention should be al-

tered and to identify which intervention option is best for whom. These variables usually include information that is useful in detecting early signs that the intervention is insufficiently effective (e.g., early signs of nonresponse to intervention, adherence, side effects, and burden), but it can also include contextual information (e.g., individual, family, social, and environmental characteristics) as well as information concerning the intervention options already received. The logic is that the best intervention option for patients varies according to different values of the tailoring variables. The fourth ingredient is a sequence of decision rules, one rule per critical decision. The decision rule links individuals? characteristics and ongoing performance with specific intervention options. The aim of these decision rules is to guide practitioners in deciding which intervention options to use at each stage of the adaptive intervention based on available information relating to the characteristics and/or ongoing performance of the patient. Each decision rule inputs the tailoring variables and outputs one or more recommended intervention options [18, 44, 45, 46]

A Just-In-Time Adaptive Intervention (JITAI) is an adaptive intervention designed to address the dynamically changing needs/behavior of patients in real-time. Compared to AI, a JITAI is more flexible in terms of the timing and location of the adaptation and delivery of intervention. While an AI usually consists of no more than 10 total decision points, the total number of decision points in a JITAI may range from 100 to 1000. While the adaptation and delivery of AI usually take place at a doctor's appointment, JITAI adapts and assigns interventions as users go about their daily life. JITAI consists of all four key elements mentioned in the last paragraph. For more details regarding an organizing framework for guiding the construction of JITAIs, refer to [58].

- **Decision points** The total number of decision points in a JITAI can be much larger. In addition, decision points in a JITAI may be selected by the scientists or specified by the user. Scientists may choose decision points at fixed time points of the day, any time when the user is at high risk of falling back to his/her unhealthy behavior. In addition, the user may request help/intervention and thus select a decision point at his/her own need.

- **Tailoring variables** Tailoring variables in a JITAI are obtained via active sensing and passive sensing. Active sensing is reported by the user through a questionnaire. It can be initiated by the user, or by the mobile devices. While Active sensing requires user engagement, passive sensing, on the contrary, uses advanced technology to assess user's environmental and social context while requiring minimal or no engagement from the user. Examples of passive sensing include GPS and accelerometers. The former is used to measure the user's geographical location and the latter is used to

measure the user's physical activity level.

- **Intervention options** While intervention options in AI is usually designed to target long-term health outcomes, intervention options in a JITAI is usually short in their duration and are targeted for behavioral change in a the moment as opposed to longer term outcome. Examples of intervention options in JITAI include encouraging messages and recommendations that target behavioral changes in a short duration followed the intervention.

- **decision rules** Similar to AI, a JITAI utilizes a sequence of decision rules, or policies, that inputs tailoring variables and outputs an intervention option.

As a concrete example, [38] have recently designed a mobile intervention, called HeartSteps, seeking to reduce users' sedentary behavior and increase physical activity such as walking and running. Installed on Android smartphones, this application is paired with Jawbone wristband to monitor users' activity data such as the total step counts everyday as well as the distribution of steps count across different location and time of the day. Heartsteps can also access users' current location, weather conditions, time of the day and day of the week. Heartsteps contains two intervention components: daily activity planning and suggestion for physical activity. When a user receives a suggestion for physical activity, s/he can respond by pressing the "thumbs-up" or "thumbs-down" buttons to indicate whether or not s/he liked the suggestion. The user also has an option to "snooze" which indicates that s/he does not want to receive any suggestions following the next 2, 4, 8, 12 hours. Decision points for Heartsteps can be anytime during the day when the smartphone is turned on with internet access. Potential tailoring variables include weather, user's activity level during the past day/week, the frequency that a user thumbs up or thumbs down, etc. Intervention options, as described, include daily activity planning and suggestion for physical activity. A policy utilizes tailoring variables to recommend appropriate interventions. An example policy is to suggest the user to walk outside for 10 minutes if the weather is sunny; otherwise suggest to user to stand up and stretch for 10 minutes.

1.2 A Review on Bandit and Contextual Bandit Algorithm

The seminal paper by Robbins [69] set the stage for an important class of sequential decision making problem, now widely known as multi-arm bandit problems. A multi-armed bandit problem is a sequential decision making problem defined by a set of actions. At each decision point, the decision maker chooses an action and observe a reward, a feedback for the action s/he has taken, before the next decision point. S/he does not get to

observed feedbacks associated with other actions; in other words, the feedback is partial. The goal of the decision maker is to maximize the his/her cumulative rewards. The multi-armed bandit problem is *stochastic* if the rewards for each action are distributed according to fixed probability distribution depending on the action and nothing else. For a stochastic multi-armed bandit problem, the quality of a decision making algorithm is measured by the expected cumulative rewards, or equivalently the expected *regret*. Regret the difference in cumulative rewards from between the algorithm and the optimality where one always choose the action with the highest expected reward. Let K denote the number of arms and $R_{i,t}$ be the random reward from pulling arm i at decision point t. Use $\{I_t\}_{t=1}^{T}$ to denote the sequence of arms that the algorithm has taken up to time T. Expected regret is the difference between the expected cumulative rewards had the decision maker always chosen the arm with the highest expected reward and the expected cumulative reward under a particular the algorithm:

$$\begin{aligned} regret &= \sum_{t=1}^{T} \max_{i:1\leq i\leq k} \mathbb{E}[R_{i,t}] - \mathbb{E}\sum_{t=1}^{T}[R_{I_t,t}] \\ &= T \max_{i:1\leq i\leq k} \mathbb{E}[R_{i,1}] - \mathbb{E}\sum_{t=1}^{T}[R_{I_t,t}] \end{aligned}$$

The most fundamental issue in tackling a multi-armed bandit problem is dealing with the *exploration and exploitation tradeoff*. Exploitation encourages pulling the seemingly best arm while exploration encourages sampling in the uncharted territory to identify the underlying best arm with high precision. Over exploitation and under exploration is associated with higher risk of being trapped at a sub-optimal arm, which inflates the regret. Under exploitation and over exploration, on the other hand, also increases the regret by sampling the sub-optimal arms with higher frequency than needed. A successful bandit algorithm is usually designed to carefully balance exploration and exploitation.

Several genres of multi-armed bandit algorithm have been proposed. [43] followed by [3] proposed the well-known Upper Confidence Bound (UCB) algorithm, for which they have proved theoretical optimal bound for the regret. UCB algorithm, at each decision point, chooses the arm with the highest upper confidence bound. Arms that have been sampled with low frequency have wider confidence bound and may be selected even if they don't have the highest estimated mean reward. Another genre of algorithms is probability matching, among which Thompson Sampling is the most popular algorithm. Using Bayesian heuristics, the invention of Thompson sampling dated back to early 1930s [82].

The idea underlying Thompson sampling has been rediscovered later, [91, 75]. The basic idea is to impose a prior distribution on the underlying parameters of the reward distribution. The algorithm updates the posterior at each decision point and select arms according to the posterior probability of being the best arm. See [13] for a comprehensive review on multi-armed bandits.

However, traditional multi-armed bandit problem are too restrictive under many circumstances. Quite often, decision makers observe side information to assist decision making. The side information may further influence the reward together with the choice of action. A generalization to multi-armed bandit was first proposed by [90] where a covariate that affects the rewards for each arm is introduced. This formulation is now widely known as *contextual bandit*. In the literature, contextual bandits are also called bandits with covariate, bandits with side information, associative bandits, and associative reinforcement learning. At decision point t, the decision maker observes a context S_t and takes an action A_t accordingly. A reward R_t, depending on both the action and the side information, is revealed before the next decision point. In contextual bandit problems, the regret is the difference in expected cumulative reward from between an contextual bandit algorithm and the optimality where one always chooses the best arm at a given context:

$$regret = \sum_{t=1}^{T} R_{A_t^*,t} - \sum_{t=1}^{T} R_{I_t,t}$$

where I_t is the algorithm-chosen arm at decision point t and $A_t^* = \operatorname{argmax} \mathbb{E}(R|S = S_t, A = a)$ is the best arm, the arm with the highest expected reward given context S_t. Contextual bandits have many applications such as online advertising, personalized news article recommendation. For example, the goal of online advertising is to display an appropriate and interesting advertisement when users visit the website to maximize the click-through-rate. The set of actions are the set of advertisement for display. Choice of the advertisement should be based on contextual information including users' previous browsing history, IP address, and other relevant information available to the advertiser.

In the following we review two of the most popular contextual bandit algorithms, both of which imposes a linear reward structure. That is, the expected reward $\mathbb{E}(R|S, A)$ is a linear function of a context-action feature vector. [92, 67, 61] have work on contextual bandit that ventures outside of the linear reward structure.

LinUCB LinUCB was introduced by [50] to extend the well known upper confidence bound (UCB) algorithm for multi-arm bandit problems to contextual bandit problems. This algorithm assumes that the expected reward is a linear function of some context-action

feature $f(S, A)$. The linear function depends on an unknown reward/weight parameter. LinUCB estimates the reward parameter at each decision point and constructs confidence interval for this parameter. When a context S_t is revealed, LinUCB calculates an upper confidence bound for the expected reward $\mathbb{E}(R|S_t, A = a)$ for all possibilities of actions. LinUCB then chooses the action that is associated with the highest upper confidence bound for S_t. LinUCB uses a tuning parameter α to control the tradeoff between exploration and exploitation: small values of α favor exploitation while larger values of α favor exploration. [16] provides theoretical justification for a master algorithm SupLinUCB that calls LinUCB as a subroutine: if SupLinUCB is run with $\alpha = \sqrt{\frac{1}{2}\ln(\frac{2TK}{\delta})}$, then with probability at least $1 - \delta$, the regret of SupLinUCB is bounded by $O(\sqrt{Td\ln^3(KT\ln(T)/\delta)})$. α is defined in algorithm 1, the implementation of LinUCB.

Algorithm 1: LinUCB

Input: A context-action feature vector $f(s, a)$ with length d. T total number of decision points. A tuning parameter $\alpha > 0$. A constant ζ to guarantee the invertibility of matrix $B(t)$

Initialization: $B(0) = \zeta I_d$, where I_d is a $d \times d$ identity matrix. $A(0) = 0_d$ is a $d \times 1$ column vector.

Start from $t = 0$.

while $t \leq T$ **do**

At decision point t, observe context S_t ;

$\hat{\mu}_t = B(t)^{-1}A(t)$;

for *a=1,.., K* **do**

$$u_{t,a} = \hat{\mu}_t^T f(S_t, a) + \alpha\sqrt{f(S_t, a)^T B(t-1)^{-1} f(S_t, a)}$$

end

Draw an action $a_t = \text{argmax}_a\, u_{t,a}$;

Observe an immediate reward R_t ;

update:

$B(t) = B(t-1) + f(S_t, A_t)f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t)R_t$;

Go to decision point $t + 1$.

end

Thompson sampling Under the same linear expected reward structure with an assumption that the error terms are R-sub-Gaussian, [1] proposed a Thompson sampling algorithm for contextual bandit and proved theoretical guarantee of the algorithm in terms of the regret bound. Thompson sampling is an old heuristic and dates back to the work of Thompson

in 1920s. The idea of Thompson sampling is choose each action by its probability of being optimal. To apply on contextual bandit problems, the algorithm starts off with a prior distribution on the reward parameter and updates the posterior at every decision point. The algorithm then calculates the posterior probability for each arm to be optimal and draws an action accordingly. The authors showed that, with probability $1-\delta$, the total regret for Thompson Sampling in time T is bounded as $\mathcal{T} = O(d^{3/2}\sqrt{T}(\ln(T) + \sqrt{\ln T \ln(1/\delta)}))$. Algorithm 2 shows the implementation of Thompson sampling contextual bandit algorithm.

Algorithm 2: The Thompson Sampling Algorithm

Input: T is the total number of decision points. A constant $0 < \delta < 1$.

$\sigma = R\sqrt{9d\ln(T/\delta)}$

Initialization: $B(0) = \zeta I_d$, where I_d is a $d \times d$ identity matrix. $A(0) = 0_d$ is a $d \times 1$ column vector.

Start from $t = 0$.

while $t \leq T$ **do**

At decision point t, observe context S_t ;

$\hat{\mu}_t = B(t)^{-1}A(t)$;

Draw $\mu \sim N(\hat{\mu}_t, \sigma^2 B(t)^{-1})$;

Choose action $a_t = \text{argmax}_a\, f(S_t, a)^T\mu$;

Observe a reward R_t ;

update:

$B(t) = B(t-1) + f(S_t, A_t)f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t)R_t$;

Go to decision point $t+1$.

end

CHAPTER 2

Online Learning of Optimal Policy: Formulation, Algorithm and Theory

In this chapter, we first formulate the online learning of optimal policy as contextual bandit problem and provide justification for doing so. We then introduce the definition of optimality: the optimal policy maximizes the average reward subject to a stochasticity constraint. By imposing a stochasticity constraint the optimal policy is stochastic, which lowers the risk of users' habituation and disengagement. Furthermore, stochasticity allows the algorithm to sufficiently explore different actions, a crucial requirement towards efficient online learning. We propose an online actor critic algorithm that learns the optimal policy. The critic imposes and estimates a linear model on the expected reward while the actor estimates the optimal policy utilizing the estimated reward parameters from the critic. Finally we develop asymptotic theory for the actor critic algorithm.

2.1 Problem formulation

2.1.1 Modeling the Decision Making Problem as a Contextual Bandit Problem

We formulate the online learning of optimal policy as a *stochastic* contextual bandit problem. Following the notation in section 1.2, a contextual bandit problem is specified by a quadruple $(\mathcal{S}, P, \mathcal{A}, \mathcal{R})$, where $\mathcal{S}$ is the context space, P is the probability law on the context space, $\mathcal{A}$ is the action space and $\mathcal{R}$ is a expected reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that maps a context-action pair to a real-valued expected reward. At decision point t, the decision maker observes a context S_t, take an action A_t after which a reward R_t is revealed before the next decision point. The decision maker does not get to observe the reward associated other actions. Contexts are i.i.d. with distribution P. Up to decision point t, the decision maker observes the data as a sequence of tuples $\{(S_\tau, A_\tau, R_\tau)\}_{\tau=1}^{t}$.

One of the strongest assumptions in contextual bandit, if not the strongest assumption, is that action A_t has a momentary effect on the reward R_t, but does not affect the distribution of S_τ for $\tau \geq t+1$. Under this assumption, one can be completely myopic and ignore the effect of an action on the distant future in searching for a good policy. In the following, we provide justification to formulate the online learning of optimal policy in mobile health as a contextual bandit problem.

The assumption that previous actions do not influence future contexts makes contextual bandit problem a simplified special case of Markov Decision Processes (MDPs). Following the notation used in [34], a MDP is specified by a 5-tuple $M = \{\mathcal{S}, \mathcal{A}, T, R, \gamma_{eval}\}$, where $\mathcal{S}$ is the context space and $\mathcal{A}$ is the decision space. $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function that specifies the probability $P(S_{t+1}|S_t, A_t)$. γ_{eval} is a discount factor that reflects how the decision maker trades off short term and long term reward. R is the expected reward function.The goal of the decision maker is to maximize the expected value of the sum of rewards discounted by γ_{eval}. A decision rule, or policy π is a mapping from the context space $\mathcal{S}$ to the decision space $\mathcal{A}$, or a probability distribution on the decision space in the case of stochastic policy. The expected utility of policy π is

$$V^{\pi}_{M,\gamma_{eval}} = \mathbb{E}(\sum_{t=0}^{\infty} \gamma^t_{eval} R_t) \tag{2.1}$$

where the expectation is taken over the randomness in context distribution, policy and realized rewards. The optimal policy is the policy that maximizes the expected utility. γ_{eval} is called *evaluation horizon*, a parameter specified by the decision maker when formulating the problem. While criterion 2.1 is the one and only criterion to evaluate the performance of a policy, a learning algorithm may use choose a *planning horizon* γ_{plan} different from the evaluation horizon. In particular, by formulating the online optimal policy learning as a contextual bandit problem and running a online contextual bandit algorithm, one essential sets $\gamma_{plan} = 0$. One may question the validity of such a choice or whether we should model the decision making problem as a full-blown MDP and use a larger γ_{plan}. The reason comes in three folds. We first justify the contextual bandit formulation is reasonable given the nature of mobile health application. We then justify the advantages of modeling and solving the problem as a contextual bandit problem even when the underlying dynamics is a MDP.

First of all, we envision that the assumption that previous actions do not influence future contexts is reasonable in many mobile health applications. Compared with other exogenous factors in users' personal lives and professional lives, often mobile health interventions have minimal and momentary effects on the contexts. Let's consider the example of Heart-

steps application introduced in the previous chapter. Contextual variables in this mobile health application includes weather, time of the day, day of time, how hectic is the user, GPS location and e.t.c. These variable are mostly influenced by events happening in users' real lives and are influenced by the physical activity suggestion from Heartsteps to a minimal extent. Contextual bandits have found applications and successes in online advertising [78], article recommendation [50] and publishing messages on social networks [42].

Second of all, solving a MDP is much more computational demanding than solving a contextual bandit problem. Contextual information in a lot of mobile health applications is highly private. For this reason, we expect a lot of the computation to be done locally on the mobile devices. Since extensive computation load reduces the battery life, this concern deems the contextual bandit formulation more appropriate. The discount factor γ_{plan} used in any planning algorithms is strongly related to the computational expense. The large γ_{plan} is, or the longer the planning horizon is, the higher the computational burden is in general [40, 35]. By choosing planning horizon shorter than the evaluation horizon, one trades off the optimality of the learned policy for computational efficiency.

Last but not least, as [34] has pointed out, choosing a short planning horizon when the transition probability is estimate from data avoids overfitting. In particular, they showed that the planning horizon γ_{plan} controls the complexity of the policy class. Choosing a planning horizon close to the evaluation horizon increases the complexity of the policy class. When the MDP model is estimated based on a finite dataset, increased complexity of policy class is associated with higher risk of overfitting. Choosing a shorter planning horizon has a regularization effect in reducing overfitting. Their reasoning is analogous to the standard bias-variance trade off argument in machine learning literature. The planning horizon γ_{plan} serves as a regularization parameter of a learning algorithm. The larger the sample size is in estimating MDP, the higher the planning horizon should be. In this dissertation, we consider the mobile health studies where the sample size is small to moderate. We therefore feel appropriate to choose the largest regularization by formulating the problem as a contextual bandit problem, .

Having been first proposed in the 70s of the last century, the contextual bandit problem has resurged in the past decades with application to online advertisement, online article recommendation, etc. However, the application to personalized intervention on mobile devices distincts our problem formulation from those in existing computer science literature. A successful contextual bandit algorithm that makes online article recommendations is usually designed to maximize certain measure of online performance, for example, the average click-through-rate [50]. The algorithm developer is less concerned about solving the mystery of which contextual variables are useful for decision making: contextual information

often sits in the web browser and is inexpensive to collect. In contrast, collecting contextual information for mobile health decision making can be expensive, time-consuming and burdensome. For example, tracking users' location using GPS on smartphones reduces battery life and may undermine overall users' experience, collecting self-reported measures on users' preferences is burdensome and may lead to intervention attrition. Since evaluating the usefulness of contexts are important in building high quality treatment policy for mobile health,

1. The policy should be easily interpretable in the sense that there is a weight associated with each component of the context to reflect how it influences the choice of intervention.

2. Scientists should be able to capture the uncertainty in the estimated weight and create confidence intervals for the weights. The confidence intervals can be used to decide whether a weight is significantly different from 0, thus answering the question whether a particular component of the context is useful for personalizing intervention.

To this end, we consider a class of stochastic parametrized treatment policies, each one of which is a mapping from the context space $\mathcal{S}$ to a probability distribution on the action space $\mathcal{A}$. In this dissertation, we consider a contextual bandit problem with a binary decision space $\mathcal{A} = \{0, 1\}$. The probability of taking action a given context s is given by a class of logistic functions:

$$\pi_\theta(A = 1|S = s) = \frac{e^{g(s)^T\theta}}{1 + e^{g(s)^T\theta}}$$
$$\pi_\theta(A = 0|S = s) = \frac{1}{1 + e^{g(s)^T\theta}}$$

where $g(s)$ is a p dimensional vector that contains candidate tailoring variables. In the parametrized policy $\pi_\theta(a|s)$ the influence of the contextual variables on the choice of action is reflected by the signs and magnitudes of θ. Statistical inferences such as confidence intervals and hypothesis testing on the optimal θ can answer the scientific question whether a particular contextual variable is useful for decision making.

2.1.2 The Regularized Average Reward

In this section, we discuss definition of optimality in learning for the optimal treatment policy for mobile health. The most natural criterion to measure the quality of a treat-

ment policy is the average reward. However, as lemma 1 will show, policy that maximizes the average reward is often deterministic. Deterministic policies may lead to treatment habituation due to the predictability and lack of variation in the recommended treatment [66, 24, 23]. To encourage intervention variety, we impose a stochasticity constraint 2.7 that requires with high probability in contexts, the optimal policy to explore all actions with a decision-maker-specified probability. The optimal policy is defined to the maximizer of a regularized average reward, the Lagrangian function of the constraint maximization problem. The optimal policy is thus guaranteed to be stochastic. A nice by-product of imposing a stochasticity constraint is that it automatically guarantees exploration. Therefore an online learning algorithm need not have an explicit exploration component such as ϵ-greedy or Boltzmann exploration.

A natural and intuitive definition of optimal policy is the policy that maximizes the average reward. For example, in developing treatment policy to promote physical activity, the objective is to increase the average daily step count. Average reward is approximated by the discounted reward 2.1 by letting γ_{eval} approach 1. In a contextual bandit formulation where contexts distribution are independent of treatment policy, the average reward of a policy $\pi_\theta(A = a|S = s)$ is the expected reward $E(R|S = s, A = a)$ weighted by the distribution of the contexts and the distribution of the action:

$$V^*(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} \mathbb{E}(R|s, a)\pi_\theta(a|s) \tag{2.2}$$

Although our focus is a class of parametrized stochastic policies, the policy that maximizes the average reward 2.2 is often deterministic. The following lemma shows that, in a simple setting where the context space is one-dimensional and finite, the policy that maximizes the average reward may be a deterministic policy.

Lemma 1. *Suppose the context space is discrete and finite, $\mathcal{S} = \{s_1, s_2, ..., s_K\}$. Among the policy class $\pi_\theta(A = 1|S = s) = \frac{e^{\theta_0+\theta_1 s}}{1+e^{\theta_0+\theta_1 s}}$, there exists a policy for which both θ_0 and θ_1 are infinite that maximizes $V^*(\theta)$. In other words, at least one of the optimal policy is a deterministic policy.*

Proof. Without the loss of generality, we assume that $0 < s_1 < s_2 < ... < s_K$. Otherwise, if some s_i's are negative, we can transform all the contexts to be positive by adding to s_i's a constant greater than $\min_{1\le i\le K} s_i$. Denote this constant by M and the corresponding policy parameter by $\tilde{\theta}$. There is a one-to-one correspondence between the two policy classes:

$$\tilde{\theta}_0 = \theta_0 - M\theta_1$$
$$\tilde{\theta}_1 = \theta_1$$

Therefore if the lemma holds when all contexts are positive the same conclusion hold in the general setting. We use $p(\theta)$ to denote the probability the probability of choosing action $A = 1$ for policy π_θ at the K different values of context:

$$(\frac{e^{\theta_0+\theta_1 s_1}}{1+e^{\theta_0+\theta_1 s_1}}, \frac{e^{\theta_0+\theta_1 s_2}}{1+e^{\theta_0+\theta_1 s_2}}, ..., \frac{e^{\theta_0+\theta_1 s_K}}{1+e^{\theta_0+\theta_1 s_K}})$$

Notice that each entry in $p(\theta)$ is number between 0 and 1 with equality if the policy is deterministic at certain context. A key step towards proving deterministic optimal policy is to show the following closed convex hull equivalency:

$$conv(\{p(\theta) : \theta \in \mathbb{R}^2\}) = conv(\{(\nu_1, ..., \nu_K), \nu_i \in \{0, 1\}, \nu_1 \leq ... \leq \nu_K \text{ or } \nu_1 \geq ... \geq \nu_K\})$$

We examine the limiting points of $p(\theta)$ when θ_0 and θ_1 tends to infinity. We consider the case where $\theta_0 \neq 0$ and let $\theta_1 = p\theta_0$ where p is a fixed value. It holds that

$$\frac{e^{\theta_0+\theta_1 s}}{1+e^{\theta_0+\theta_1 s}} = \frac{e^{\theta_0(1+ps)}}{1+e^{\theta_0(1+ps)}} \rightarrow \begin{cases} 0 : if\theta_0 \rightarrow -\infty, p > -1/s \\ 0 : if\theta_0 \rightarrow \infty, p < -1/s \\ 1 : if\theta_0 \rightarrow -\infty, p < -1/s \\ 1 : if\theta_0 \rightarrow \infty, p > -1/s \end{cases}$$

It follows that when $\theta_0 \rightarrow -\infty$ and p scans through the $K+1$ intervals on $\mathbb{R}$: $(-\infty, -1/s_1]$, $(-1/s_1, -1/s_2]$, ... $(-1/s_K, \infty)$, $p(\theta)$ approaches the following $K + 1$ limiting points:

$$(1, 1, ..., 1)$$
$$(0, 1, ..., 1)$$
$$...$$
$$(0, 0, ..., 1)$$
$$(0, 0, ..., 0)$$

when $\theta_0 \rightarrow \infty$ and p scans through the $K + 1$ intervals, $p(\theta)$ approaches the following $K + 1$ limiting points

$$(0,0,...,0)$$
$$(1,0,...,0)$$
$$...$$
$$(1,1,...,0)$$
$$(1,1,...,1)$$

There are in total $2K$ limiting points: $\{(\nu_1,...,\nu_K), \nu_i \in \{0,1\}, \nu_1 \leq ... \leq \nu_K$ or $\nu_1 \geq ... \geq \nu_K\}$. Each limiting point is a K dimensional vector with 0-1 entries in an either increasing or decreasing order. Now we show that any $p(\theta), \theta \in \mathbb{R}^2$ is a convex combination of the limiting points. Let $p(\theta) = [p_1(\theta), p_2(\theta), ..., p_K(\theta)]$. In fact,

- If $\theta_1 = 0$, $p(\theta) = (1 - p_1(\theta))(0,0,...,0) + p_1(\theta)(1,1,...,1)$
- If $\theta_1 > 0$, we have $0 < p_1(\theta) < p_2(\theta) < ... < p_K(\theta) < 1$ and

$$\begin{aligned} p(\theta) = p_1(\theta)(1,1,...,1) + (p_2(\theta) - p_1(\theta))(0,1,...,1) + ... \\ +(p_K(\theta) - p_{K-1}(\theta))(0,0,...,1) + (1 - p_K(\theta)) * (0,0,...,0) \end{aligned}$$

- If $\theta_1 < 0$, we have $1 > p_1(\theta) > p_2(\theta) > ... > p_K(\theta) > 0$ and

$$\begin{aligned} p(\theta) = (1 - p_1(\theta)) * (0,0,...,0) + (p_1(\theta) - p_2(\theta))(1,0,...,0) + ... \\ +(p_K(\theta) - p_{K-1}(\theta))(1,1,...,0) + p_K(\theta)(1,1,...,1) \end{aligned}$$

Returning to optimizing the average reward, we denote $\alpha_i = P(S = s_i)(\mathbb{E}(R|S = s_i, A = 1) - \mathbb{E}(R|S = s_i, A = 0))$.

$$\max_{\theta} V^*(\theta) = \max_{\theta} \sum_{i=1}^{K} \alpha_i p_i(\theta) \tag{2.3}$$

$$= \max_{(p_1,\ldots,p_K)\in\{p(\theta):\theta\in\mathbb{R}^2\}} \sum_{i=1}^{K} \alpha_i p_i \tag{2.4}$$

$$= \max_{(p_1,\ldots,p_K)\in conv(\{p(\theta):\theta\in\mathbb{R}^2\})} \sum_{i=1}^{K} \alpha_i p_i \tag{2.5}$$

$$= \max_{(p_1,\ldots,p_K)\in conv(\{(\nu_1,\ldots,\nu_K),\nu_i\in\{0,1\},\nu_1\leq\ldots\leq\nu_K \text{ or } \nu_1\geq\ldots\geq\nu_K\})} \sum_{i=1}^{K} \alpha_i p_i \tag{2.6}$$

. Equation from 2.4 to 2.5 is followed by the fact that the objective function is linear (and thus convex) in p_i's. Equivalency from 2.5 to 2.6 is a direct product of the closed convex hull equivalency. Theories in linear programming theory suggests that one of the maximal points is attained at the vertices of the convex hull of the feasible set. Therefore we have proved that one of the policy that maximizes $V^*(\theta)$ is deterministic. □

Behavioral science literature has documented many empirical evidences and theory that deterministic treatment policies lead to habituation and that intervention variety has proven to be therapeutic by preventing/retarding habituation [66, 24, 23]. To encourage intervention variety, we make sure that the treatment policies sufficiently explores all available actions. When the action space is binary, which is the focus of this article, one way to mathematize intervention variety is to introduce a chance constraint [88, 79] that with high probability, probability taken with respect to the context, the probability of taking each action is bounded away from 0:

$$P_s(p_0 \leq \pi_\theta(A = 1|S) \leq 1 - p_0) \geq 1 - \alpha \tag{2.7}$$

where $0 < p_0 < 0.5$, $0 < \alpha < 1$ are scientists-specified constants controlling the amount of stochasticity. P_s the probability law over the context space. The stochasticity constraint requires that, for at least $(1-\alpha)100\%$ of the contexts, there is at least p_0 probability to take both actions.

Maximizing the average reward $V^*(\theta)$ subject to the stochasticity constraint 2.7 is a chance constrained optimization problem, an active research area in recent years [59, 15]. Solving this chance constraint problem, however, involves a major difficulty – constraint 2.7 is in general a non-convex constraint on θ for many possible distribution of the context

and many possible forms of the constraint function. Moreover, the left hand side of the stochasticity constraint is an expectation of an indicator function. Both the non-convexity and the non-smoothness of this inequality make the optimization problem computationally intractable. We circumvent this difficulty by replacing constraint 2.7 with a convex alternative. By applying the Markov inequality, we reach a relaxed and smoother stochasticity constraint that produces computational tractability:

$$\theta^T \mathbb{E}[g(S)^T g(S)]\theta \leq (\log(\frac{p_0}{1-p_0}))^2 \alpha \tag{2.8}$$

The quadratic constraint is more stringent than the stochasticity constraint and always guarantees *at least* the amount of intervention variety the scientists have asked for. We define the *optimal policy* to be the policy θ that maximizes the average reward $V^*(\theta)$ subject to the quadratic constraint 2.8, i.e., the maximum of the following quadratic constrained optimization problem:

$$\max_{\theta} V^*(\theta), \text{ s. t. } \theta^T \mathbb{E}[g(S)^T g(S)]\theta \leq (\log(\frac{p_0}{1-p_0}))^2 \alpha \tag{2.9}$$

Instead of solving the quadratic optimization problem, we maximize the corresponding Lagrangian function. Incorporating inequality constraints by forming Lagrangian has been widely used in reinforcement learning literature to solve constrained Markov decision problem [12, 8]. Given a Lagrangian multiplier λ, the following Lagrangian function $J^*_\lambda(\theta)$:

$$J^*_\lambda(\theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} E(R|S=s, A=a)\pi_\theta(s,a) - \lambda \theta^T \mathbb{E}[g(S)^T g(S)]\theta \tag{2.10}$$

is referred to as the *regularized average reward* in this article. The optimal policy is the maximizer of the regularized average reward:

$$\theta^* = \operatorname*{argmax}_{\theta} J^*_\lambda(\theta) \tag{2.11}$$

There are two computational advantages to maximize the regularized average reward. One advantage is that optimizing the regularized average reward function remains a well-defined optimization problem even when there is no treatment effect. When the expected reward does not depend on the action, $\mathbb{E}(R|S=s, A=a) = \mathbb{E}(R|S=s)$, constrained maximization of the average reward $V^*(\theta)$ is an ill-posed problem due to the lack of unique

solution. In fact, all policies in the feasible set have the same average reward. The regularized average reward function, in contrast, has a unique maximizer at $\theta = 0_p$, a pure random policy that assigns both actions with 50% probability. Therefore, maximizing the regularized average reward gives rise to a unique and sensible solution when there is no treatment effect. The other advantage to maximize the regularized average reward function is computational. When the uniqueness of optimal policy is not an issue, maximization of $J^*_\lambda(\theta)$ has computational advantages over maximization of $V^*(\theta)$ under constraint 2.8 because the subtraction of the quadratic term $\lambda\theta^T\mathbb{E}[g(S)^T g(S)]\theta$ introduces concavity to the surface of $J^*_\lambda(\theta)$, thus speeding up the computation.

A natural question to ask, when transforming the constrained optimization problem 2.9 to an unconstrained one 2.11, is whether a Lagrangian multiplier exists for each level of stringency of the quadratic constraint. While the correspondence between the constrained optimization and the unconstrained one may not seem so obvious due to the lack of convexity in $V^*(\theta)$, we established the following lemma 2 given assumption 1. Assumption 1 assumes the uniqueness of the global maximum for all positive λ. While assumption 1 seems strong and hard to verify analytically, we have verified that this assumption holds in our numerical experiment in chapter 3. In assumption 2 we assume the positive definiteness of the matrix $\mathbb{E}(g(S)g(S)^T)$.

Assumption 1. *For every $0 < \lambda < \infty$, the global maximum of $J^*_\lambda(\theta)$ is a singleton.*

Assumption 2. *Positive Definiteness: The matrix $\mathbb{E}(g(S)g(S)^T)$ is positive definite.*

Lemma 2. *If the maximizer of the average reward function $V^*(\theta)$ is deterministic, i.e. $P(\pi_\theta(A = 1|S) = 1) > 0$ or $P(\pi_\theta(A = 0|S) = 1) > 0$, under assumption 1 and 2, for every $K = (\log(\frac{p_0}{1-p_0}))^2\alpha > 0$ there exist a $\lambda > 0$ such that the solution of the constrained optimization problem 2.9 is the solution of the unconstrained optimization problem 2.11.*

Proof. Let θ^*_λ be one of the global maxima of the Lagrangian function: $\theta^*_\lambda = \text{argmax}_\theta\, J^*_\lambda(\theta)$. Let $\beta_\lambda = \theta^{*T}_\lambda\mathbb{E}[g(S)^T g(S)]\theta^*_\lambda$. By proposition 3.3.4 in [7], θ^*_λ is a global maximum of constrained problem:

$$\begin{aligned}&\max_\theta V^*(\theta)\\ &\text{s.t. } \theta^T\mathbb{E}[g(S)^T g(S)]\theta \le \beta_\lambda\end{aligned}$$

In addition, the stringency of the quadratic constraint increases monotonically with the value of the Lagrangian coefficient λ. Let $0 < \lambda_1 < \lambda_2$ and with some abuse of notation, let θ_1 and θ_2 be (one of) the global maximals of Lagrangian function $J^*_{\lambda_1}(\theta)$ and $J^*_{\lambda_2}(\theta)$. It follows that

$$
\begin{aligned}
&- V^*(\theta_2) + \lambda_2 \theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2 \\
\leq &- V^*(\theta_1) + \lambda_2 \theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 \\
= &- V^*(\theta_1) + \lambda_1 \theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 + (\lambda_2 - \lambda_1)\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 \\
\leq &- V^*(\theta_2) + \lambda_1 \theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2 + (\lambda_2 - \lambda_1)\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1
\end{aligned}
$$

It follows that

$$
\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 \geq \theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2
$$

. As λ approaches 0, the maximal of the regularized average reward approaches the maximal of the average reward function, for which $\mathbb{E}(\theta^T g(S))^2 \to \infty$. As λ increases towards ∞, maximal of the regularized average reward approaches the random policy with $\theta = 0$. It's only left to show that $\theta_\lambda^{*T} \mathbb{E}[g(S)^T g(S)]\theta_\lambda^*$ is a continuous function of λ. Under assumption 1, we can verify that conditions in Theorem 2.2 in [25] holds. This theorem implies that the solution set of the unconstrained optimization 2.11 is continuous in λ, sufficient to conclude the continuity of $\theta_\lambda^{*T} \mathbb{E}[g(S)^T g(S)]\theta_\lambda^*$. □

2.2 An Online Actor Critic Algorithm

In this section, we propose an online actor critic algorithm for the learning of optimal policy parameter. The idea of actor critic originates from the literature of reinforcement learning [41, 9, 84]. There, as the maximizer of the long-term discounted/average reward, the optimal policy parameter is updated iteratively using stochastic gradient descent where the gradient depends on the Q-function, the long-term discounted/average reward given a particular state-action pair [77]. The learning algorithm is decomposed into two step: in critic step the algorithm estimates the Q-function or value function after which the algorithm uses the estimated Q-function or value function to update the policy. Actor critic algorithms to solve MDP is usually two-time scaled where the actor updates at a slower rate than the critic. The reason to do so is that both the stationary distribution of states and the Q-function or value function depend on the policy.

Likewise, in a contextual bandit problem, estimation of the optimal policy requires assistance from estimation of the expected reward $\mathbb{E}(R|S, A)$. The observed "training data" at decision point t is a stream of triples $\{S_\tau, A_\tau, R_\tau\}_{\tau=1}^t$. The optimal policy can be estimated by maximizing the aforementioned regularized average reward, which can be estimated

empirically by:

$$J_\lambda(\theta) = \frac{1}{t}\sum_{\tau=1}^{t}\sum_{a}\mathbb{E}(R|S_\tau, a)\pi_\theta(A=a|S_\tau) - \lambda\theta^T[\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T]\theta \tag{2.12}$$

which requires the knowledge of $\mathbb{E}(R|S, A)$. In the proposed actor critic algorithm, the critic estimates the expected reward; the estimated expected is then plugged into 2.12 to produce an estimated optimal policy. The estimated optimal policy is used to select an action at the next decision point.

2.2.1 The Critic with a Linear Function Approximation

We use $\mathbb{E}(R|S=s, A=a) = R(s,a)$ to denote the expected reward given context s and action a. We make the following two assumptions regarding the expected reward.

Assumption 3. *Linear realizability assumption: given context $S=s$ and action $A=a$, the reward has expectation $E(R|S=s, A=a) = f(s,a)^T\mu^*$ plus a noise variable ϵ with sub-Gaussian distribution. The noise variables at different decision points are i.i.d. with mean 0 and variance σ^2.*

This assumption is often referred to as the "linear realizability assumption" in existing contextual multi-arm bandits literature and is a standard assumption in this literature [1, 2, 26, 50, 16]. In addition, given context S_t and action A_t, the difference between the realized reward R_t and the expected reward $R(S_t, A_t)$ is $\epsilon_t = R_t - R(S_t, A_t)$. ϵ_t are assumed to be i.i.d with mean 0 and have finite second moment.

Assumption 4. *The error terms in the reward model are i.i.d with mean 0 and have finite second moment.*

The reward parameter μ is estimated by the ordinary least square estimator [1]

$$\hat{\mu}_t = (\sum_{\tau=1}^{t} f(S_\tau, A_\tau)f(S_\tau, A_\tau)^T)^{-1}\sum_{\tau=1}^{t} f(S_\tau, A_\tau)R_\tau \tag{2.13}$$

. Compared to the usual ordinary least square estimation, the reward features 2.13 are non-i.i.id. Action A_τ is drawn according to the estimated optimal policy at decision point

[1] When running the critic algorithm online, however, the matrix $\sum_{\tau=1}^{t} f(s_\tau, a_\tau)f(s_\tau, a_\tau)^T$ may not have full rank when t is small. For this reason, we introduce a small regularization term when calculating the least square estimate. See the initialization of $B(t)$ in algorithm 1.

$\tau - 1$, which depends on the entire history at or before decision point $\tau - 1$. The dependency introduced presents challenges in analyzing the actor critic algorithm. Details will be presented in section 2.3.

2.2.2 The Actor and the Actor Critic Algorithm

Once an estimated reward parameter is obtained from the critic, the actor estimates the optimal policy parameter by maximizing the estimated average reward. With some abuse of notation, we denote the regularized average reward function under the reward parameter μ by

$$J_\lambda(\theta, \mu) = \sum_s d(s) \sum_a f(s,a)^T \mu \pi_\theta(a|s) - \lambda \theta^T g(s) g(s)^T \theta \tag{2.14}$$

where $d(s)$ is the stationary distribution of the context. In other words $J_\lambda(\theta, \mu^*) = J_\lambda^*(\theta)$. Plugging $\hat{\mu}_t$ into display 2.12, an estimate to the regularized average reward function at decision point t is

$$\hat{J}_t(\theta, \hat{\mu}_t) = \mathbb{P}_t j(\hat{\mu}_t, \theta, S) \tag{2.15}$$

$$= \frac{1}{t} \sum_{\tau=1}^{t} \sum_a f(S_\tau, a)^T \hat{\mu}_t \pi_\theta(A = a|S_\tau) - \lambda \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_i)^T g(S_i)] \theta \tag{2.16}$$

where $\mathbb{P}$ is the empirical probability law on $\mathcal{S} \times \mathcal{A}$. The estimated optimal policy parameter is

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmax}}\, \hat{J}_t(\theta, \hat{\mu}_t) \tag{2.17}$$

We propose an actor-critic on linear learning algorithm to learn the optimal treatment policy as described in Algorithm 3. Inputs of the actor critic algorithm includes, the total number of decision points, T which is usually determined by intervention duration in number of days and the frequency of the decision points in a single day. Inputs of the algorithm also includes specifying the reward feature $f(s,a)$ and the policy feature $g(s)$. The former can be chosen using model selection techniques on historical dataset. The latter consists of candidate tailoring variable, usually specified by domain scientists. Matrix $B(t)$ is used to store the summation of the outer product of reward features; $B(0)$ is ini-

tialized to be an identity matrix multiplied by a small coefficient ζ, because the matrix $\sum_{\tau=1}^{t} f(S_\tau, A_\tau) f(S_\tau, A_\tau)^T$ may not have full rank when t is small. $A(t)$ is used to store the summation of $f(S_t, A_t) R_t$; $A(0)$ is initialized to be a d dimensional column matrix with all zero entries. The initial treatment policy θ_0 is chosen to be the domain knowledge driven policy, or based on historical data if available. At each decision point, the algorithm acquires a new context S_t, takes an action according to policy $\pi_{\theta_{t-1}}$ and then observes a reward R_t before the next decision point. The critic updates the reward parameter according to 2.13; the actor updates the optimal policy parameter according to 2.17.

Algorithm 3: An online actor critic algorithm with linear expected reward

Input of algorithm: T, the total number of decision points; reward feature $f(s, a)$ with dimension d; policy feature $g(s)$ with dimension p.

Critic initialization: $B(0) = \zeta I_{d\times d}$, a $d \times d$ identity matrix. $A(0) = 0_d$ is a $d \times 1$ column vector.

Actor initialization: θ_0 is the best treatment policy based on domain theory of historical data.

Start from $t = 0$.

while $t \leq T$ **do**

At decision point t, observe context S_t ;

Draw an action a_t according to probability distribution $\pi_{\theta_{t-1}}(A|S_t)$;

Observe an immediate reward R_t ;

Critic update:

$B(t) = B(t-1) + f(S_t, A_t) f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t) R_t$,

$\hat{\mu}_t = B(t)^{-1} A(t)$. ;

Actor update:

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} f(S_\tau, a)^T \mu_t \pi_\theta(A = a|S_\tau) - \lambda \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_i)^T g(S_i)] \theta \tag{2.18}$$

Go to decision point $t + 1$.

end

2.3 Asymptotic Theory of the Actor Critic Algorithm

In this section, we analyze the consistency and the asymptotic normality of the actor critic algorithm. We first show, in theorem 1 and theorem 2 respectively, that the estimated re-

ward parameter and the estimated optimal policy parameter converge in probability to their population counterparts as the number of decision points increases. We analyze the asymptotic normality of the estimated reward parameter and estimated optimal policy parameter in theorem 3 and 4. In addition to the aforementioned assumptions, we make the following assumptions:

Assumption 5. *Boundedness: The reward R, reward feature $f(S, A)$ and the reward parameter μ^* are bounded with probability one. Without loss of generality, we assume that $|\mu^*|_2 < 1$ and $|f(S, A)|_2 \leq 1$ with probability one.*

Assumption 6. *Positive Definiteness: The matrix $\mathbb{E}(g(S)g(S)^T) = \sum_{s\in\mathcal{S}} d(s)g(s)g(s)^T$ is positive definite.*

As the very first step towards establishing the asymptotic properties of the actor critic algorithm, we show that, for a fixed Lagrangian multiplier λ the optimal policy parameter that maximizes the regularized average reward 2.10 essentially lies in a bounded set. Moreover, the estimated optimal policy parameter is bounded with probability going to 1. Lemma 3 sets the foundation for us to leverage the existing statistical asymptotic theories.

Lemma 3. *Assume that assumption 5 and 6 holds. Given a fixed λ, the population optimal policy θ^* lies in a compact set. In addition, the estimated optimal policy $\hat{\theta}_t$ lies in a compact set with probability going to 1.*

Proof. This lemma is proved by comparing the regularized average reward function $J^*_\lambda(\theta)$ at θ^* and at 0_p. The optimal regularized average reward is:

$$\begin{aligned} J^*_\lambda(\theta^*) = & \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} f(s,a)^T \mu^* \pi_{\theta^*}(A = a|S = s) - \lambda\theta^{*T}\mathbb{E}[g(S)^T g(S)]\theta^* \\ \leq & \sum_{s,a} d(s) \frac{|f(s,a)|_2^2 + |\mu^*|_2^2}{2} \pi_{\theta^*}(A = a|S = s) - \lambda\theta^{*T}\mathbb{E}[g(S)^T g(S)]\theta^* \\ \leq & \quad 1 - \lambda\theta^{*T}\mathbb{E}[g(S)^T g(S)]\theta^* \end{aligned}$$

While on the other hand the regularized average reward for the random policy $\theta = 0_p$ is

$$J^*_\lambda(0_p) = \sum_{s,a} d(s) f(s,a)^T \mu^*/2 \geq 0$$

By the optimality of policy θ^*, $1 - \lambda\theta^T\mathbb{E}[g(S)^T g(S)]\theta \geq 0$, which leads to the necessary condition for the optimal policy parameter:

$$\theta^{*T}\mathbb{E}[g(S)^T g(S)]\theta^* \leq \frac{1}{\lambda} \tag{2.19}$$

According to assumption 6, the above inequality defines a bounded ellipsoid for θ^*, which concludes the first part of lemma 3. To prove the second conclusion of Lemma 3, we notice that $\hat{\mu}_t$ is bounded since the smallest eigenvalue of $B(t)$ is bounded away from 0 by ζ and both the reward feature and the reward is bounded with probability 1. Denote the bound by K. By comparing $\hat{J}_t(\theta, \hat{\mu}_t)$ at $\theta = \hat{\theta}_t$ and $\theta = 0_p$ we have

$$\hat{\theta}_t^T[\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T]\hat{\theta}_t \leq \frac{K}{\lambda} \tag{2.20}$$

It remains to show that the smallest eigenvalue of $\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T$ is bounded away from 0 with probability going to 1. Using the matrix Chernoff inequality, theorem 1 in [83], for any $0 < \delta < 1$,

$$P\{\lambda_{min}(\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T) \leq (1-\delta)\lambda_{min}(\mathbb{E}g(S)g(S)^T)\} \leq p[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}]^{\frac{t\lambda_{min}(\mathbb{E}g(S)g(S)^T)}{R}} \tag{2.21}$$

where R is the maximal eigenvalue of $g(S)g(S)^T$ and p is the dimension of $g(S)$. Taking $\delta = 0.5$, the right-hand side of the Chernoff inequality goes to 0 as t goes to ∞. Therefore with probability going to 1, inequality 2.20 defines a compact set on $\mathbb{R}^p$. We have proved the second part of the lemma.

□

Following this lemma, we make the following assumption:

Assumption 7. *The matrix* $\mathbb{E}_\theta(f(S,A)f(S,A)^T) = \sum_{s\in\mathcal{S}} d(s)\sum_a f(s,a)f(s,a)^T\pi_\theta(A = a|S = a)$ *is positive definite for* θ *in the compact set in Lemma 3.*

Theorem 1. *Consistency of the critic: Under assumptions 3 through 7, the critic's estimate* $\hat{\mu}_t$ *converges to the true reward parameter* μ^* *in probability.*

The non-i.i.d. nature of sample presents challenges in developing the asymptotic theory. In particular, A_t depends the entire trajectory of observations before decision point t as well as the context at the current decision point. The challenges in proving the consistency of the reward parameter estimation is solved by exploiting the closed form of $\hat{\mu}_t$ and applying the

matrix Azuma's inequality, Theorem 7.1 in [83]. We notice that, in proving the consistency of $\hat{\mu}_t$, that the $\hat{\mu}_t$ is consistent as long as the data generating policy $\hat{\theta}_t$ lies in a compact set with probability going to 1, which guarantees a minimum exploration probability. The proof of theorem 1 is outlined by showing that $\frac{B(t)^{-1}}{t}$ is bounded with probability going to one and that $\frac{A(t)}{t}$ converges to 0 in probability. Details can be found in the appendix.

One of the most critical assumptions in deriving the consistency of M-estimator is the uniqueness of global maximum of the criterion function. Let $h(\theta) = \mathbb{E}_X(H(\theta, X))$ be the population level criterion function and θ^* be the global maxima. It is often assumed that given any constant $\delta > 0$, there exists a neighborhood of θ^*, denoted by $B(\theta^*, \epsilon)$ where ϵ measures the "diameter" of the neighborhood, such that $h(\theta)$ is bounded above by $h(\theta^*) - \delta$ for θ outside the neighborhood. The estimated optimal policy parameter $\hat{\theta}_t$ is "almost" an M-estimator except that the empirical criterion function depends not only on the empirical distribution of context but also the estimated reward parameter $\hat{\mu}_t$. We therefore make this assumption uniform in a neighborhood of μ^*:

Assumption 8. *Uniform separateness of the global maximum: There exists a neighborhood of μ^* such that the following holds. $J(\theta, \mu)$ as a function of θ has unique global maximum for all μ in this neighborhood of μ^*. Moreover, for any $\delta > 0$, there exists $\epsilon > 0$ such that*

$$J(\theta^{\mu}, \mu) - \max_{\theta \notin B(\theta^{\mu}, \epsilon)} J(\theta, \mu) \geq \delta \tag{2.22}$$

for all μ in this particular neighborhood of μ^*.

Under the aforementioned assumptions, the following theorem states the consistency of the estimation of optimal policy parameter.

Theorem 2. *Consistency of the actor: Under assumption 3 through assumption 8, the actor's estimate $\hat{\theta}_t$ converges to true optimal policy parameter θ^* in probability.*

The two steps in proving theorem 2 are to first show that if the sequence $\hat{\mu}_t$ converges to μ^*, then $\tilde{\theta}_t = \text{argmax}_\theta \, J(\theta, \mu_t)$, the optimal policy based on the population distribution of context, converges to θ^*, where

$$J(\mu, \theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \mathbb{E}(R|S = s, A = a)\pi_\theta(A = a|S = s) - \lambda \theta^T \mathbb{E}[g(S)g(S)^T]\theta \tag{2.23}$$

We then show that $\theta_t^{\mu} = \text{argmax}_\theta \, \hat{J}_t(\theta, \mu)$ converges to $\theta^{\mu} = \text{argmax}_\theta \, J(\theta, \mu)$ uniformly in a neighborhood of μ^*. Details of the proof can be found in the appendix.

Theorem 3 states the asymptotic normality of the critic. Proof of theorem 3 relies on the vector-valued central limit theorem, the details of which can be found in the appendix.

Theorem 3. *Asymptotic normality of the critic: Under assumption 3 through assumption 7,* $\sqrt{t}(\hat{\mu}_t - \mu^*)$ *converges in distribution to multivariate normal with mean* 0_d *and covariance matrix*
$[\mathbb{E}_{\theta^*}(f(S,A)f(S,A)^T)]^{-1}\sigma^2$*, where*

$$\mathbb{E}_\theta(f(S,A)f(S,A)^T) = \sum_s d(s) \sum_a f(s,a)f(s,a)^T \pi_\theta(s,a)$$

is the expected value of $f(S,A)f(S,A)^T$ *under policy* θ*.* σ *is the standard deviation of* ϵ_t*. The plug-in estimator of the asymptotic covariance is consistent.*

The asymptotic normality of $\hat{\theta}_t$ is established based on the asymptotic normality of $\hat{\mu}_t$ and that the class of random functions $\{j(\theta, \mu, S) : \theta \in \mathbb{R}^p, |\mu|_2 \leq 1\}$ are P-Donsker. $j(\theta, \mu, S)$ is the expected reward for context S under policy θ and reward parameter μ. See details of the proof in the appendix.

Theorem 4. *Asymptotic normality of the actor: Under assumption 3 through assumption 8,* $\sqrt{t}(\hat{\theta}_t - \theta^*)$ *converges in distribution to multivariate normal with mean* 0_p *and sandwich covariance matrix*

$$[J_{\theta\theta}(\mu^*, \theta^*]^{-1} V^* [J_{\theta\theta}(\mu^*, \theta^*)]^{-1} \tag{2.24}$$

, where $V^* = \sigma^2 J_{\theta\mu}(\mu^*, \theta^*)\mathbb{E}_\theta(f(S,A)f(S,A)^T J_{\mu\theta}(\mu^*, \theta^*) + \mathbb{E}[j_\theta(\mu^*, \theta^*, S) j_\theta(\mu^*, \theta^*, S)^T]$*. In the expression of asymptotic covariance matrix,*

$$j(\mu, \theta, S) = \sum_a f(S,a)^T \mu \pi_\theta(A = a|S) - \lambda \theta^T [g(S)g(S)^T]\theta \tag{2.25}$$

and $J(\mu, \theta)$ *is defined in 2.23.* $J_{\theta\theta}$ *and* $J_{\theta\mu}$ *are the second order partial derivatives of* J *with respect to* θ *twice and with respect* θ *and* μ*, respectively.* j_θ *is the first order partial derivative of* j *with respect to* θ*. The following plug-in estimator of the asymptotic covariance is consistent.*

$$\begin{aligned}(\hat{J}_{\theta\theta}(\mu^*, \theta^*)^{-1} [\hat{\sigma}^2 \hat{J}_{\theta\mu}(\mu^*, \theta^*) \hat{\mathbb{E}}_\theta(f(s,a)f(s,a)^T) \hat{J}_{\mu\theta}(\mu^*, \theta^*) \\ + \frac{1}{t} \sum_{i=1}^{t} j_\theta(\mu^*, \theta^*, s_i) j_\theta(\mu^*, \theta^*, s_i)^T] (\hat{J}_{\theta\theta}(\mu^*, \theta^*))^{-1}\end{aligned} \tag{2.26}$$

A bound on the expected regret can be derived as a by-product of the square-root convergence rate of $\hat{\theta}_t$. The expected regret of an online algorithm is the difference between the expected reward under the algorithm and that under the optimal policy θ^*:

$$\text{expected regret} = T \sum_s d(s) \sum_a \mathbb{E}(R|S=a, A=a)\pi_{\theta^*}(S=s|A=a) - \mathbb{E}[\sum_{t=1}^{T} R_t] \tag{2.27}$$

where $\{R_t\}_{t=1}^T$ is the sequence of rewards generated in the algorithm. Straightforward calculation shows that

$$\begin{aligned}\text{expected regret} &= \sum_{t=1}^{T} \sum_s d(s) \sum_a \mathbb{E}(R|S=a, A=a)[\pi_{\theta^*}(S=s|A=a) - \mathbb{E}(\pi_{\hat{\theta}_{t-1}}(S=s|A=a))] \\ &= \sum_{t=1}^{T} \sum_s d(s) \sum_a \mathbb{E}(R|S=a, A=a)\mathbb{E}(\pi'_{\hat{\theta}_{t,s,a}}(S=s|A=a)(\theta^* - \hat{\theta}_{t-1}))\end{aligned}$$

where $\hat{\theta}_{t,s,a}$ is a random variable that lands between θ^* and $\hat{\theta}_{t-1}$. Under the boundedness assumption on the expected reward, Theorem 4 implies the following Corollary:

Corollary 1. *The expected regret of the actor critic algorithm 3 is of order* $O(\sqrt{T})$.

We shall point that the regret bound provided in the corollary is not comparable to the regret bound derived for LinUCB and Thompson Sampling where there is no assumption on the distribution of the contexts.

2.4 Small Sample Variance estimation and Bootstrap Confidence intervals

In this section, we discuss issues, challenges and solutions in creating confidence intervals for the optimal policy parameter θ^* when the sample size, the total number of decision points, is small. We use a simple example to illustrate that the traditional plug-in variance estimator is plagued with underestimation issue, the direct consequence of which is the deflated confidence levels of the Wald-type confidence intervals for θ^*. We propose to use bootstrap confidence intervals when the sample size is finite. Evaluation of the bootstrap confidence intervals will be provided in chapter 3.

2.4.1 Plug-in Variance Estimation and Wald Confidence intervals

One of the most straightforward ways to estimate the asymptotic variance of θ_t is through the plug-in variance estimation, the formulae of which is provided in Theorem 4. Once an estimated variance $\hat{V}_i$ is obtained for $\sqrt{t}(\hat{\theta}_i - \theta_i^*)$, a $(1-2\alpha)\%$ Wald type confidence interval for θ_i^* has the form: $[\hat{\theta}_i - z_\alpha \frac{\hat{V}_i}{\sqrt{t}}, \hat{\theta}_i + z_\alpha \frac{\hat{V}_i}{\sqrt{t}}]$. Here θ_i is the i-th component in θ and z_α is the upper 100α percentile of a standard normal distribution. The plug-in variance estimator and the associated Wald confidence intervals work well in many statistics problems. We shall see that, however, the plug-in variance estimator of the estimated optimal policy parameters suffers from underestimation issue in small to moderate sample sizes. In particular this estimator is very sensitive to the plugged-in value of the estimated reward parameter and policy parameter: a small deviation from the true parameters can result in an inflated or deflated variance estimation. Deflated variance estimation produces anti-conservative confidence intervals, a grossly undesirable property for confidence intervals. The following simple example illustrates the problem.

Example 1. *The context is binary with probability distribution* $\mathbb{P}(S=1) = \mathbb{P}(S=-1) = 0.5$. *The reward is generated according to the following linear model: given context* $S \in \{-1, 1\}$ *and action* $A \in \{0, 1\}$,

$$R = \mu_0^* + \mu_1^* S + \mu_2^* A + \mu_3^* SA + \epsilon$$

where ϵ *follows a normal distribution with mean zero and standard deviation 9. The true reward parameter is* $\mu^* = [1, 1, 1, 1]$. *Both* μ^* *and the standard deviation of* ϵ *are chosen to approximate the realistic signal noise ratio in mobile health applications. We consider the policy class* $\pi_\theta(A=1|S=s) = \frac{e^{\theta_0+\theta_1 s}}{1+e^{\theta_0+\theta_1 s}}$.

The differences between the plug-in estimated variance and its population counterpart are that (1) the former uses the empirical distribution of context to replace the unknown population distribution and (2) the unknown reward parameter and optimal policy parameter are replaced by their estimates. We emphasis that it is the second difference that leads to the underestimated variance in small sample size. To see this, we ignore the difference between the empirical distribution and the population distribution of contexts, which is very small for sample size $T \geq 50$ *under a Bernoulli context distribution with equal probability. Now the plug-in variance estimator is a function of the estimated reward parameter* $\hat{\mu}_t$ *and the estimated policy parameter* $\hat{\theta}_t$. *Notice that in 2.17,* $\hat{\theta}_t = [\hat{\theta}_{t,0}, \hat{\theta}_{t,1}]$ *is a function of* $\hat{\mu}_t = [\hat{\mu}_{t,0}, \hat{\mu}_{t,1}, \hat{\mu}_{t,2}, \hat{\mu}_{t,3}]$ *and the empirical distribution of context. If we replace the empirical distribution in calculating* $\hat{\theta}_t$ *by its population counterpart,* $\hat{\theta}_t$ *is simply a function of* $\hat{\mu}_t$. *In the rest part of the example, we drop the subscript* t *in the estimated reward*

parameter and denote the estimate of μ_2 and μ_3 by $\hat{\mu}_2$ and $\hat{\mu}_3$, respectively. Likewise, $\hat{\theta}_{t,i}$ is replaced by $\hat{\theta}_i$ for $i = 0, 1$.

Figure 2.1 is the surface plot showing how the plug-in variance estimation changes as function of the estimated reward parameter. The surface plot of the plug-in variance estimation has a mountain-like pattern with two ridges along the two diagonals $\hat{\mu}_2 + \hat{\mu}_3 = 0$ and $\hat{\mu}_2 - \hat{\mu}_3 = 0$. The height of the ridge increases as both $\hat{\mu}_2$ and $\hat{\mu}_3$ approaches the origin. The peak of mountain is at the origin where $\hat{\mu}_2 = \hat{\mu}_3 = 0$. The true reward parameter $(\mu_2^*, \mu_3^*) = (1, 1)$ is close to the origin and lies right on the one of the ridges. There are four "valleys" where the combinations of $\hat{\mu}_2$ and $\hat{\mu}_3$ gives a small plug-in variance. The fluctuation in the plug-in variance estimator can be roughly explained by the curvature of the estimated regularized average reward function:

- *When $\hat{\mu}_2 = \hat{\mu}_3 = 0$, $J(\theta, \hat{\mu}) = -\lambda\|\theta\|_2^2 + \sum_s d(s)(\hat{\mu}_0 + \hat{\mu}_1 s)$. The curvature of J as a function of θ is completely determined by the Lagrangian term $-\lambda\|\theta\|_2^2$.*

- *When $\hat{\mu}_2 = \hat{\mu}_3$, $J(\theta, \hat{\mu}) = (\hat{\mu}_2 + \hat{\mu}_3)\pi_\theta(A = 1|S = 1) - \lambda\|\theta\|_2^2 + \sum_s d(s)(\hat{\mu}_0 + \hat{\mu}_1 s)$. The curvature of J is contributed by the two terms $(\hat{\mu}_2 + \hat{\mu}_3)\pi_\theta(A = 1|S = 1)$ and the Lagrangian term.*

- *When $\hat{\mu}_2 = -\hat{\mu}_3$, $J(\theta, \hat{\mu}) = (\hat{\mu}_2 - \hat{\mu}_3)\pi_\theta(A = 1|S = -1) - \lambda\|\theta\|_2^2 + \sum_s d(s)(\hat{\mu}_0 + \hat{\mu}_1 s)$. The curvature of J is contributed by the two terms $(\hat{\mu}_2 - \hat{\mu}_3)\pi_\theta(A = 1|S = -1)$ and the Lagrangian term.*

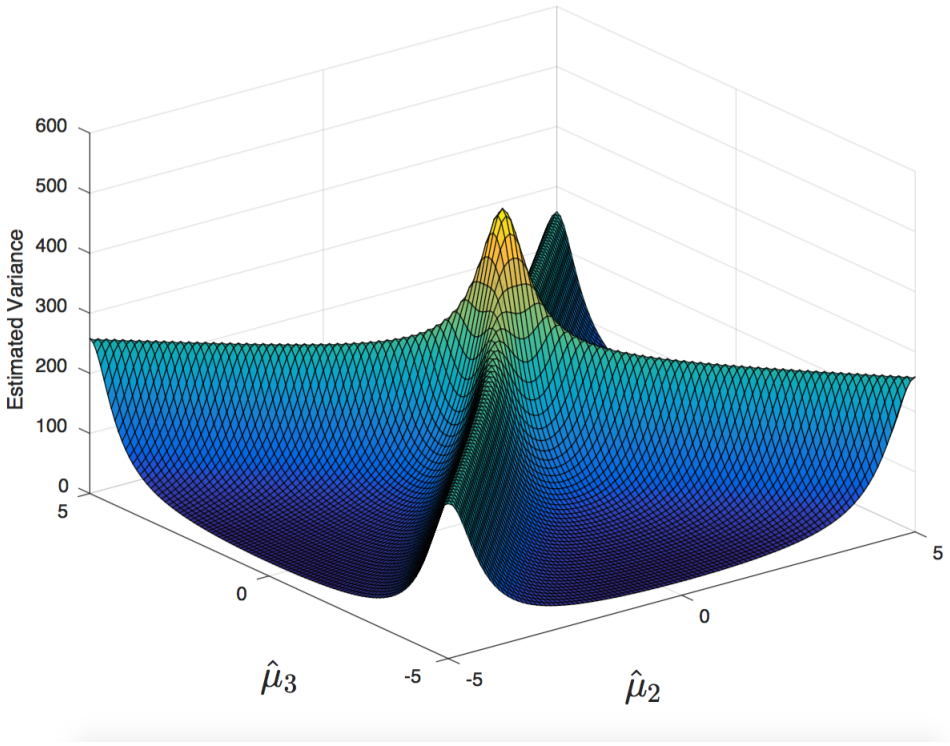

Figure 2.1: Plug in variance estimation as a function of
$\hat{\mu}_2$ and $\hat{\mu}_3$, x axis represents $\hat{\mu}_{t,2}$, y axis represents $\hat{\mu}_{t,3}$ and z axis represents the plug-in asymptotic variance of $\hat{\theta}_0$ with $\lambda = 0.1$

Due to large areas of valley the plug-in variance estimation is biased down, a direct consequence of which is the anti-conservatism of the Wald confidence intervals. We perform a simulation study using the toy generative model described above. The simulation consists of 1000 repetitions of running the online actor critic algorithm and recording the end-of-study statistics, including the plugin variance estimate, the Wald confidence intervals and the theoretical Wald confidence intervals based on the true asymptotic variance. The first two columns in table 2.1 show the bias of plug-in variance at different sample sizes. At all three different sample sizes, the plug-in variance estimator underestimates the true asymptotic variance, which is 293.03 *for both policy parameters. Column 3 and column 4 show the coverage rate of the Wald-type confidence interval (CI) using the plug-in estimated variance. It is not surprising that the confidence intervals suffer from severe anti-conservatism, a consequence of the heavily biased variance estimation. Column 5 and 6 show the coverage rate of the Wald-type confidence interval based on the true asymptotic variance. Comparing the coverage rates, it is clear that the anti-conservatism is due to the underestimated variance.*

sample size	bias in variance estimation		coverage of Wald CI (%)		coverage of theoretical Wald CI (%)	
	θ_0	θ_1	θ_0	θ_1	θ_0	θ_1
100	-181.56	-181.56	75.5	74.9	100.0	100.0
250	-131.71	-131.71	77.9	77.3	98.5	98.1
500	-108.64	-108.64	78.8	79.2	98.9	98.7

Table 2.1: Underestimation of the plug-in variance estimator and the Wald confidence intervals. Theoretical Wald CI is created based on the true asymptotic variance.

To detail how the confidence interval coverage is connected with the estimated reward parameter $(\hat{\mu}_2, \hat{\mu}_3)$, figure 2.2 and figure 2.3 present two scatter plots of $\hat{\mu}_2, \hat{\mu}_3$ for the 1000 simulated datasets at sample size 100 and 500. Different colors are used to mark the datasets where the confidence intervals of both θ_0 and θ_1 cover the true parameter (blue), only one of them cover the truth (green), neither of them covers the truth (fading yellow). The true parameter are marked with a red asterisk. Indeed the yellow points and green points are in the "valleys". Some of the blue points are away from truth, but nevertheless they remain on the ridge, which produces a high variance estimate. Comparing the two scatter plots, as the sample size increases, the estimated reward parameter is less spread out. Nevertheless there are still significantly many pair of $\hat{\mu}_2, \hat{\mu}_3$ that fall in the "valleys", leading to a underestimated variance and anti-conservative confidence intervals.

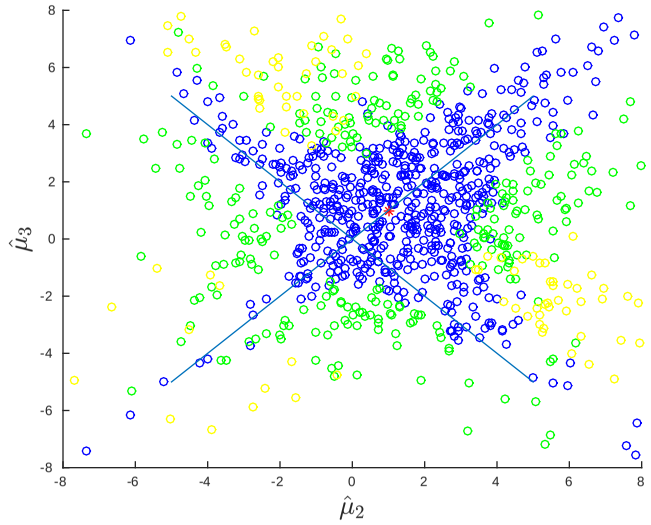

Figure 2.2: Wald confidence interval coverage for 1000 simulated datasets as a function of $\hat{\mu}_3$ and $\hat{\mu}_2$ at sample size 100.

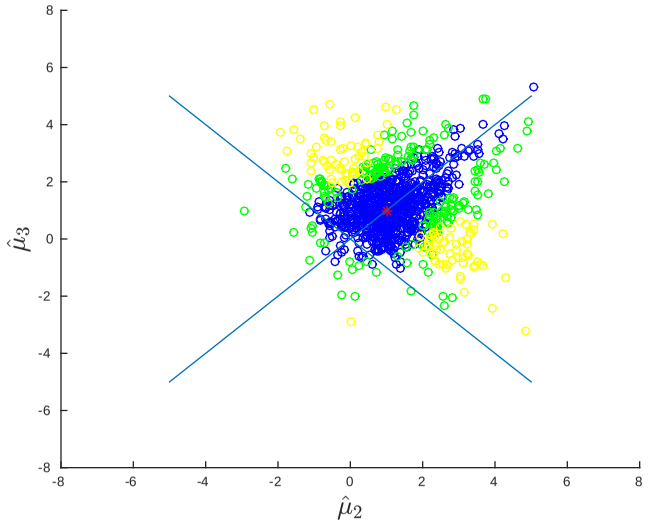

Figure 2.3: Wald confidence interval coverage in 1000 simulated datasets as a function of $\hat{\mu}_3$ and $\hat{\mu}_2$ at sample size 500.

Figure 2.4 shows the histogram for the normalized distance $\frac{\sqrt{\hat{T}}(\theta_i - \theta_i^*)}{\hat{V}_i}$ *for* $i = 0, 1$ *where* $T = 100$*. This is the distance between the estimated and the true optimal policy parameter normalized by the estimated asymptotic variance. For the Wald confidence intervals to have descent coverage, histogram of the normalized distances need to approximate a standard normal distribution. However, as figure 2.4 suggests, the histograms have heavier tails compared to a standard normal due to the underestimated variance. The figure also suggests that the percentile-t bootstrap confidence intervals can be a good remedy.*

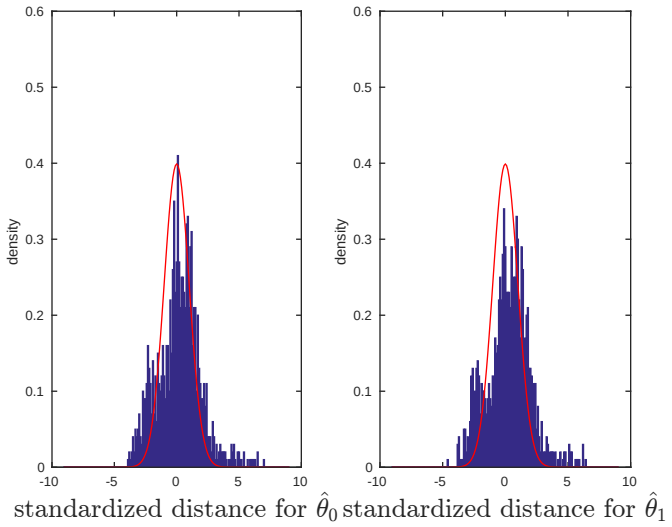

Figure 2.4: Histograms of the normalized distance $\frac{\sqrt{\hat{T}}(\theta_i - \theta_i^*)}{\hat{V}_i}$ for $i = 0, 1$ at sample size 100

2.4.2 Bootstrap Confidence intervals

Our solution to the anti-conservative Wald confidence interval is the bootstrap confidence interval. Upon completion of the online actor critic algorithm with a total number of T decision points, we have recorded a sequence of contexts $\{S_i\}_{i=1}^T$ and rewards $\{R_i\}_{i=1}^T$. We also have the estimated reward parameter $\hat{\mu}_T$ and optimal policy $\hat{\theta}_T$ estimated at the very last decision point. The sample of reward noise is created by $\{\epsilon_t = R_t - f(S_t, A_t)^T \mu_T\}_{t=1}^T$. We obtain a bootstrap sample for the estimated optimal policy $\hat{\theta}_T^b$ as described in algorithm 4. In generating the bootstrapped samples, the sequence of contexts are fixed both in their values and order as $\{S_i\}_{i=1}^T$. At each decision point, the algorithm chooses an action based on the estimated optimal policy from the previous decision point. A bootstrapped residual is then generated by sampling without replace from $\{\epsilon_t\}_{t=1}^T$ to create a bootstrapped reward R_t^b. The critic and the actor then update their estimates just like algorithm 3. At the exit of algorithm 4 (at decision point T), we obtain a pair of $\hat{\mu}_T^b$ and $\hat{\theta}_T^b$. We use the pair

to obtain a plug-in variance estimate $\hat{V}^b$. Repeating algorithm 4 for a total of B times to get a bootstrap sample of the estimated optimal policy $\{\hat{\theta}_T^b\}_{b=1}^B$ and plug-in variance estimates $\{\hat{V}_T^b\}_{b=1}^B$. We create bootstrap percentile-t confidence intervals for θ_i^*, the i-th component of the optimal policy parameter. For each θ_i^*, we use the empirical percentile of $\{\frac{\sqrt{t}(\hat{\theta}_{T,i}^b - \hat{\theta}_{T,i})}{\sqrt{\hat{V}_T^b}}\}_{b=1}^B$, denoted by p_α to replace the normal distribution percentile in Wald confidence intervals. A $(1-2\alpha)\%$ confidence interval is

$$[\hat{\theta}_{T,i} - p_\alpha \frac{\hat{V}_i}{\sqrt{T}}, \hat{\theta}_{T,i} + p_\alpha \frac{\hat{V}_i}{\sqrt{T}}] \tag{2.28}$$

where $\hat{\theta}_{T,i}$ is the i-th component of $\hat{\theta}_T$.

Algorithm 4: Generating a bootstrap sample $\theta_T^b, \hat{V}_T^b$

Inputs: The observed context history $\{S_t\}_{t=1}^T$. A bootstrap sample of residuals $\{\epsilon_t^b\}_{t=1}^T$

Critic initialization: $B(0) = \zeta I_{d\times d}$, a $d \times d$ identity matrix. $A(0) = 0_d$ is a $d \times 1$ column vector.

Actor initialization: θ_0 is the best treatment policy based on domain theory of historical data.

while $t < T$ **do**

Context is S_t ;

Draw an action A_t^b according to policy $\pi_{\hat{\theta}_t^b}$;

Generate a bootstrap reward $R_t^b = f(S_t, A_t^b)^T \mu_T + \epsilon_t^b$;

Critic update:

$B(t) = B(t-1) + f(S_t, A_t) f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t) R_t^b$;

$\hat{\mu}_t^b = A(t)^{-1} B(t)$;

Actor update:

$$\hat{\theta}_t^b = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_a f(S_\tau, a)^T \hat{\mu}_t^b \pi_\theta(A = a|S_t) - \lambda \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau, 1)^T g(S_\tau, 1)]\theta$$

Go to decision point $t+1$;

end

Plugin μ_T^b and θ_T^b to 2.26 to get a bootstrapped variance estimate $\hat{V}^b$.

2.5 Appendix

Proof of theorem 1.

Proof. Based on the expression of $\hat{\mu}_t$, its $\mathcal{L}_2$ distance from μ^* is

$$|\hat{\mu}_t - \mu^*|^2 = A(t)B(t)^{-1}B(t)^{-1}A(t) + o_p(1) \tag{2.29}$$

$$= \frac{A(t)}{t}(\frac{B(t)}{t})^{-1}(\frac{B(t)}{t})^{-1}\frac{A(t)}{t} + o_p(1) \tag{2.30}$$

where $A(t)$ and $B(t)$ are defined in algorithm 3. The two steps to prove that $|\mu_t - \mu^*|_2^2 \to 0$ in probability are

1. $\frac{B(t)^{-1}}{t}$ is bounded with probability going to 1, and
2. $\frac{A(t)}{t}$ converges to 0 in probability.

To prove the first step, we construct a matrix-valued martingale difference sequence. Define $K(\theta) = \mathbb{E}_\theta[f(S,A)f(S,A)^T] = \sum_s d(s)\sum_a f(s,a)f(s,a)^T\pi_\theta(A=a|S=s)$

$$\begin{aligned} X_i = & \quad f(S_i,A_i)f(S_i,A_i)^T - \mathbb{E}(f(S_i,A_i)f(S_i,A_i)^T|\mathcal{F}_i) \\ = & \quad f(S_i,A_i)f(S_i,A_i)^T - \sum_s d(s)\sum_a f(s,a)f(s,a)^T\pi_{\theta_{i-1}}(a|s) \\ = & \quad f(s_i,a_i)f(s_i,a_i)^T - K(\theta_{i-1}) \end{aligned}$$

where the filtration $\mathcal{F}_i = \sigma\{\hat{\theta}_j, j \le i-1\}$ is the sigma algebra expand by the estimated optimal policy before decision point i. By assumption 5, the sequence of random matrices $\{X_i\}$ are uniformly bounded. Applying the matrix Azuma inequality in [83], it follows that

$$\lambda_{max}(\frac{B(t)}{t} - \frac{\sum_{i=1}^t K(\theta_{i-1})}{t}) \to 0 \text{ in probability}$$
$$\lambda_{min}(\frac{B(t)}{t} - \frac{\sum_{i=1}^t K(\theta_{i-1})}{t}) \to 0 \text{ in probability}$$

Let the operators λ_{min} and λ_{max} represent the smallest and the largest eigenvalue of a matrix.

$$\lambda_{min}(\frac{B(t)}{t}) = \lambda_{min}(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\theta_{i-1})}{t} + \frac{\sum_{i=1}^{t} K(\theta_{i-1})}{t})$$
$$\geq \lambda_{min}(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\theta_{i-1})}{t}) + \lambda_{min}(\frac{\sum_{i=1}^{t} K(\theta_{i-1})}{t})$$

By assumption 7, the second term $\lambda_{min}(\frac{\sum_{i=1}^{t} K(\theta_{i-1})}{t})$ is bounded with probability going to 1. Hence we have shown that the minimal eigenvalue of $\frac{B(t)}{t}$ is bounded with probability going to 1. Using the same proving techniques we can show that the maximal eigenvalue of $(\frac{B(t)}{t})^{-1}$ is bounded with probability going to 1.

The second step in proving theorem 1 is standard. Using the same filtration $\mathcal{F}_i$, we construct vector-valued martingale difference sequence $Y_i = f(S_i, A_i)\epsilon_i$. The sequence has bounded variance under assumption 5. The in-probability convergence of $\frac{A(t)}{t}$ to 0 follows immediately by applying the vector-valued Azuma inequality [33].

□

Proof of theorem 2:

Proof. Proof of the theorem consists of two steps. As the first step, we claim that if a sequence μ_t converges to μ^*, $\tilde{\theta}_t = argmax_\theta J(\theta, \mu_t)$ converges to θ^*. By Lemma 9.1 in [36], $J(\theta, \mu)$ is an absolute continuous function. We proof the claim by contradiction. Suppose the claim does not hold, i.e. there exist ϵ such that $\|\tilde{\theta}_t - \theta^*\|_2 \geq \epsilon$ for all t by taking a subsequence if necessary. The optimality of $\tilde{\theta}_t$ implies that the inequality $J(\tilde{\theta}_t, \mu_t) \geq J(\theta^*, \mu_t)$ holds for all t. Since $\tilde{\theta}_t$ is bounded, it converges to an accumulation point $\tilde{\theta}$ by taking a subsequence if necessary. Let $t \to \infty$ we have $J(\tilde{\theta}, \mu^*) \geq J(\theta^*, \mu^*)$. On the other hand $\|\theta^{**} - \theta^*\|_2 \geq \epsilon$, which contradicts with assumption 1.

As the second step, we prove that the following M estimator converges uniformly in a neighborhood of μ^*, namely

$$\theta_t^\mu = argmax_\theta \hat{J}_t(\theta, \mu) \to \theta^\mu = argmax_\theta J(\theta, \mu) \tag{2.31}$$

in probability, and uniformly over all μ in a neighborhood of μ^*. Arguments in the proof are parallel to those in Theorem 9.4 in [36]. The key is to observe that the class of random functions $\{j(\theta, \mu, s) : \theta \in \mathbb{R}^p, |\mu|_2 \leq 1\}$ are Glivenko-Cantelli. □

Proof of theorem 3

Proof.

$$\begin{aligned}\mu_t - \mu^* &= (\zeta I_d + \sum_{i=1}^{t} f(S_i, A_i) f(S_i, A_i)^T)^{-1} (\sum_{i=1}^{t} f(S_i, A_i)\epsilon_i - \mu^*) \\ &= (\frac{\zeta I_d + \sum_{i=1}^{t} f(S_i, A_i) f(S_i, A_i)^T}{t})^{-1} \sqrt{t} \frac{\sum_{i=1}^{t} f(S_i, A_i)\epsilon_i}{t} + o_p(1)\end{aligned}$$

Based on the consistency of θ_t, we have that $\frac{\zeta I_d + \sum_{i=1}^{t} f(S_i,A_i) f(S_i,A_i)^T}{t}$ converges in probability to $\mathbb{E}_{\theta^*}(f(S,A)f(S,A)^T$. Now it is the key to analyze the asymptotic distribution of the martingale difference sequence $\{f(S_i, A_i)\epsilon_i\}_{i=1}^{t}$. With respect to filtration $\mathcal{F}_{t,j} = \sigma(\{S_i, A_i, \epsilon_i\}_{i=1}^{j})$. Define $M* = [\mathbb{E}_{\theta^*}(f(S,A)f(S,A)^T)]^{-1/2}$ and a martingale difference sequence $\{\xi_{t,i} = \frac{M^* f(s_i,a_i)\epsilon_i}{\sqrt{t}}\}_{i=1}^{t}$ which is adapted to the filtration $\mathcal{F}_{t,j}$ and satisfies $\mathbb{E}(\xi_{t,i}|\mathcal{F}_{t,i-1}) = 0$, To apply vector Lindberg-Levy central limit theorem for martingale difference sequences [11], we check the two conditions in this theorem:

1. The conditional variance assumption.

$$\begin{aligned}V_t &= \sum_{i=1}^{t} \mathbb{E}(\xi_{t,i}^2 | \mathcal{F}_{t,i-1}) \\ &= \frac{1}{t} \sum_{i=1}^{t} M^* \mathbb{E}_{\theta_{i-1}}(f(s,a)f(s,a)^T) M^*\end{aligned}$$

   converges in probability to $I_d\sigma^2$ by consistency of θ_t.

2. The Lindeberg condition. For any given $\delta > 0$,

$$\begin{aligned}&\sum_{i=1}^{t} \mathbb{E}(\xi_{t,i}^2 \mathbb{I}(\|\xi_{t,i}\|_2 > \delta) | \mathcal{F}_{t,i-1}) \\ =&\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}(M^* f(S_i, A_i) f(S_i, A_i)^T \epsilon_i^2 M^* \mathbb{I}(\|M^* f(S_i, A_i)\epsilon_i\|_1 > \sqrt{t}\delta) | \mathcal{F}_{t,i-1}) \\ \leq&\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}(M^* f(S_i, A_i) f(S_i, A_i)^T \epsilon_i^2 M^* \mathbb{I}(\|M^* f(S_i, A_i)\|_2 \epsilon_i^2 > \sqrt{t}\delta) | \mathcal{F}_{t,i-1})\end{aligned}$$

   By assumption 5 $f(S, A)$ are bounded almost surely, therefore the above expression goes to 0 as $t \to 0$.

The Lindberg-Levy martingale central limit theorem concludes that

$$\sum_{i=1}^{t} \xi_{t,i} \to N(0_d, I_d\sigma^2) \text{ in distribution}$$

Therefore

$$\sqrt{t}(\hat{\mu}_t - \mu^*) \to N(0_d, [\mathbb{E}_{\theta^*}(f(S,A)f(S,A)^T)]^{-1}\sigma^2) \tag{2.32}$$

□

Proof of theorem 4.

Proof. We first prove that

$$\mathbb{G}_t j_\theta(\hat{\mu}_t, \hat{\theta}_t, S) - \mathbb{G}_t j_\theta(\mu^*, \theta^*, S) = o_p(1) \tag{2.33}$$

, where $\mathbb{G}_t = \sqrt{t}(\mathbb{P}_t - P)$, the empirical process induced by the "marginal" stochastic process $\{S_i\}_{i=1}^t$ formed by the history of contexts. The "full" stochastic process involves the sequence of triples $\{S_i, A_i, \epsilon_i\}_{i=1}^t$, the complete history of contexts, actions and reward errors. We consider the class of functions $\mathcal{F} = \{j_\theta(\mu, \theta, s) : \|\theta - \theta^*\|_2 \leq \delta, \|\mu - \mu^*\|_2 \leq \delta\}$, where $j_\theta(\mu, \theta, s)$ is the partial derivative with respect to θ of function:

$$j(\mu, \theta, s) = \sum_a f(s,a)^T \mu \pi_\theta(s,a) - \lambda \theta^T g(s) g(s)^T \theta$$

The boundedness assumption on reward feature, policy feature and reward ensures that the parametrized class of functions $j_\theta(\mu, \theta, s)$ is P-Donsker in a neighborhood of (μ^*, θ^*). In other words $\mathcal{F}$ is P-Donsker, where P is the distribution of the marginal stochastic process formed by contexts. We complete the first part of the proof by modiftying Lemma 19.24 in [85]. It may seem that the dependence of $\hat{\mu}_t$ and $\hat{\theta}_t$ on the full stochastic process could introduce complexity but a closer inspection shows that the proof goes through. The random function $j_\theta(\hat{\mu}_t, \hat{\theta}_t, S)$ belongs to the P-Donsker class defined above and satisfies that

$$\sum_s d(s)(j_\theta(\hat{\mu}_t, \hat{\theta}_t, s) - j_\theta(\mu^*, \theta^*, s))^2 \to 0$$

in probability. This is a result of the consistency of both $\hat{\mu}_t$ and $\hat{\theta}_t$, as well as apply-

ing the continuous mapping theorem. By theorem 18.10(v) in [85], $(\mathbb{G}_t, j_\theta(\hat{\mu}_t, \hat{\theta}_t, s)) \to (\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s))$ in distribution, where $\mathbb{G}_p$ is the P-Brownian bridge. The key here is that theorem 18.10 only relies on the convergence of two stochastic processes, regardlessly of whether the stochastic processes consist of i.i.d. observations and whether or not the two processes are dependent. By Lemma 18.15 in [85], almost all sample paths of $\mathbb{G}_p$ are continuous on $\mathcal{F}$. Define a mapping $h : l(\mathcal{F})^\infty \times \mathcal{F} \to \mathbb{R}$ by $h(z, f) = z(f) - z(j_\theta(\mu^*, \theta^*, s))$, which is continuous at almost every point of $(\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s))$. By the continuous mapping theorem, we have

$$\mathbb{G}_t(j_\theta(\hat{\mu}_t, \hat{\theta}_t, s) - j_\theta(\mu^*, \theta^*, s)) = h(\mathbb{G}_t, j_\theta(\hat{\mu}_t, \hat{\theta}_t, s)) \to h(\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s)) = 0$$

in distribution and thus in probability, therefore 2.33 holds.

The second part of the proof begins by noticing that θ_t satisfies the estimating equation $\mathbb{P}_t j_\theta(\hat{\mu}_t, \hat{\theta}_t, s) = 0$, so we have

$$\begin{aligned}
\mathbb{G}_t j_\theta(\hat{\mu}_t, \hat{\theta}_t, s) &= \sqrt{t}(P j_\theta(\mu^*, \theta^*, s) - P j_\theta(\hat{\mu}_t, \hat{\theta}_t, s)) \\
&= \sqrt{t}(J_\theta(\mu^*, \theta^*) - J_\theta(\hat{\mu}_t, \hat{\theta}_t)) \\
&= \sqrt{t} J^*_{\theta\theta}(\theta^* - \theta_t) + \sqrt{t} J^*_{\theta\mu}(\mu^* - \hat{\mu}_t) + \sqrt{t} o_p(\|\hat{\theta}_t - \theta^*\|) + o_p(1)
\end{aligned}$$

Together with 2.33 the above implies

$$\begin{aligned}
\sqrt{t}(\theta^* - \hat{\theta}_t) &= (J^*_{\theta\theta})^{-1} J^*_{\theta\mu} \sqrt{t}(\hat{\mu}_t - \mu^*) + \sqrt{t} o_p(\|\hat{\theta}_t - \theta^*\|) + (J^*_{\theta\theta})^{-1} \mathbb{G}_t j_\theta(\mu^*, \theta^*, s) + o_p(1) \\
&= O_p(1) + \sqrt{t} o_p(\|\hat{\theta}_t - \theta^*\|)
\end{aligned}$$

where $J^*_{\theta\theta}$ and $J^*_{\theta\mu}$ are $J_{\theta\theta}$ and $J_{\theta\mu}$ evaluated at (θ^*, μ^*). The $\sqrt{t}$ consistency of $\hat{\theta}_t$ follows through. Now 2.34 has become

$$\sqrt{t}(\theta^* - \hat{\theta}_t) = (J^*_{\theta\theta})^{-1} J^*_{\theta\mu} \sqrt{t}(\hat{\mu}_t - \mu^*) + (J^*_{\theta\theta})^{-1} \mathbb{G}_t j_\theta(\mu^*, \theta^*, S) + o_p(1) \tag{2.34}$$

Since both the two non-vanishing terms on the righthand side are asymptotically normal with zero mean, $\sqrt{t}(\theta^* - \theta_t)$ is asymptotically normal. The only task left is to derive the asymptotic variance. Plugging in the formula for $\hat{\mu}_t$, we have

$$\sqrt{t}(\theta^* - \hat{\theta}_t) = (J^*_{\theta\theta})^{-1} \frac{\sum_{i=1}^t J^*_{\theta\mu} B^* f(S_i, A_i)\epsilon_i + j_\theta(\mu^*, \theta^*, S_i)}{t} + o_p(1)$$
$$= (J^*_{\theta\theta})^{-1} \sum_{i=1}^t \zeta_{t,i} + o_p(1)$$

where $B^* = (M^*)^2 = [\mathbb{E}_{\theta^*} f(S, A) f(S, A)^T]^{-1}$. $\{\zeta_i = \frac{J^*_{\theta\mu} B^* f(S_i, A_i)\epsilon_i + j_\theta(\mu^*, \theta^*, S_i)}{t}\}_{i=1}^t$ is a martingale difference sequence with asymptotic variance

$$\sum_{i=1}^t \mathbb{E}(\zeta^2_{t,i} | \mathcal{F}_{t,i})$$
$$= \frac{1}{t} \sum_{i=1}^t \mathbb{E}(\epsilon_i^2 g^*_{\theta\mu} B^* f(S_i, A_i) f(S_i, A_i)^T B^* g^*_{\mu\theta}$$
$$+ j_\theta(\mu^*, \theta^*, S_i) j_\theta(\mu^*, \theta^*, S_i)^T - 2 J_{\theta\mu} B^* f(S_i, A_i) j_\theta(\mu^*, \theta^*, S_i)^T \epsilon_i | \mathcal{F}_{t,i})$$
$$= \frac{1}{t} \sum_{i=1}^t \sigma^2 J^*_{\theta\mu} B^* \mathbb{E}_{\theta_{i-1}}(f(S, A) f(S, A)^T) B^* J^*_{\mu\theta} + \sum_s d(s) j_\theta(\mu^*, \theta^*, s) j_\theta(\mu^*, \theta^*, s)^T$$

which converges in probability to $V^* = \sigma^2 J^*_{\theta\mu} B^* J^*_{\mu\theta} + \sum_s d(s) j_\theta(\mu^*, \theta^*, s) j_\theta(\mu^*, \theta^*, s)^T$. Therefore the asymptotic variance of $\sqrt{t}(\theta^* - \theta_t)$ is $(J^*_{\theta\theta})^{-1} V^* (J^*_{\theta\theta})^{-1}$.

□

CHAPTER 3

Numerical Experiments

In this section, we use numerical experiments to test the performance of actor-critic algorithm and the bootstrap confidence intervals proposed in the previous sections under a variety of generative models. In section 3.1, we first assess the accuracy of the estimated optimal policy parameters and the conservatism of the bootstrap confidence intervals when contexts at different decision points are i.i.d.. We expect the estimated optimal policy parameters to converge to the population optimal policy parameter as the total number of decision points T increases. In section 3.2, the context dynamics deviate from i.i.d. and are instead generated by a first degree auto regression process (AR(1)): context distribution at decision point $t + 1$ depends on the context at decision point t. We expect the performance of the algorithm and the estimated optimal policy to be pretty robust. In section 3.3.1 and 3.3.2, we create generative models that break the most crucial assumption in contextual bandit that actions do not influence future context. We allow distributions of the contexts to depend on previous actions in three different ways. In section 3.3.1, one component of the contexts is affected by previous actions through a learning effect: users pick up the skills through previous mobile interventions to maintain healthy habit. In section 3.3.2, one component of the contexts is affected by previous actions through a burden effect, which describes overly-frequent intervention tends to disengage the users. In both section 3.3.1 and 3.3.2, there is a parameter, ν for the learning effect and τ for the burden effect, that controls the size of the effect, or the amount of violation from the contextual bandit assumption that the previous actions do not influence future context. We evaluate how the performance of the contextual bandit actor critic algorithm deteriorates when the amount of violation increases.

The generative model is motivated by the Heartsteps application for improving daily physical activity [39, 20]. Heartsteps is mobile health application seeking to reduce users' sedentary behavior and increase physical activity such as walking and running. Installed on Android smartphones, this application is paired with Jawbone wristband to monitor users' activity data such as the total step counts everyday as well as the distribution of steps count

across different location and time of the day. Heartsteps can also access users' current location, weather conditions, time of the day and day of the week. Heartsteps provides suggestions for physical activity. For the purpose of testing the actor critic algorithm, our generative model foregoes some of complexities in Heartsteps application and focuses on suggestion for physical activity only. There are three decision points, appropriate time points for intervention, everyday: one in the morning, one the afternoon and one in the evening. At each decision point, Heartstep decides whether to "push" a suggestion for physical activity $A_t = 1$ or remain silent $A_t = 0$. The decision is tailored to users' current contexts. For simplicity our simulation assumes that the context at decision point t consists of three components: $S_t = [S_{t,1}, S_{t,2}, S_{t,3}]$. The three components are:

- $S_{t,1} = weather$, with $S_{t,1} = -\infty$ being extremely severe and unfriendly weather for any outdoor activities and $S_{t,1} = \infty$ being the opposite.
- $S_{t,2}$ reflects users' learning ability from previous physical activity suggestions. $S_{t,2} = \infty$ represents that the user has picked up all the skills to maintain a high level of daily physical activity while $S_{t,2} = -\infty$ represents the opposite.
- $S_{t,3}$ is a composite measure of disengagement or feeling of burden to HeartSteps application. $S_{t,3} = -\infty$ reflects an extreme state that the user is paying full attention to HeartSteps and willing to follow any its activity suggestions and $S_{t,3} = \infty$ being the opposite.

The goal of Heartstep is to reduce users' sedentary behavior. We define the cost to be the per hour sedentary time between two decision points. Cost at a decision point depends on both the previous action and the previous context. In our simulation, the cost is generated according the following linear model:

$$C_t \;=\; 10 - .4S_{t,1} - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1} + 0.2S_{t,2}) + 0.4S_{t,3} + \xi_{t,0}$$

where $\xi_{t,0}$ follows i.i.d. with standard normal distribution. In this linear model, higher values of S_1 and S_2, good weather and higher learning effect, are associated with less sedentary time while a higher value of S_3, disengagement, leads to increased sedentary time. The negative main effect of A_t indicates that physical activity suggestion ($A_t = 1$) has an treatment effect on reducing sedentary behavior compared to no suggestion $A_t = 0$. The negative interaction effects between A_t and $S_{t,1}$ and between A_t and $S_{t,2}$ reflect that physical activity suggestions are more effective when the weather condition is activity friendly or the users are equipped with good learning skills.

We study the class of parametrized policies that include all three components of context as candidate tailoring variables. The probability of recommending a physical activity suggestion is given by the following logistic function.

$$\pi_\theta(A = 1|S = [S_1, S_2, S_3]) = \frac{e^{\theta 0+\theta_1 S_1+\theta_2 S_2+\theta_3 S_3}}{1 + e^{\theta 0+\theta_1 S_1+\theta_2 S_2+\theta_3 S_3}}$$

The long term average cost under policy π_θ is:

$$C(\theta) = \sum_s d_\theta(s) \sum_a \mathbb{E}(C|S = a, A = a)\pi_\theta(A = a|S = s)$$

where $d_\theta(s)$ is the stationary distribution of context under policy π_θ. When actions have no impact on context distributions, the stationary distribution $d(s)$ does not depend on the policy parameter θ. In this case, the long term average cost reduces to the average cost:

$$C(\theta) = \sum_s d(s) \sum_a \mathbb{E}(C|S = a, A = a)\pi_\theta(A = a|S = s)$$

This is true, for example, for the types of generative model we shall investigate in section 3.1 and 3.2. The types of generative model we investigate in section 3.3.1 and 3.3.2 allow actions to impact context distributions at future decision points. There, the stationary distribution of context depends on the policy parameter θ. A stochasticity constraint specifies the proportion of contexts for which a minimal amount of exploration probability is guaranteed. As mentioned in section **, the stochasticity constraint is introduced to prevent habituation and facilitate learning.The stochasticity constraint specifies that for at least $100(1 - \alpha)\%$ context, there is at least p_0 probability of selecting both actions:

$$P\left[p_0 \leq \pi_\theta(A = 1|S_t) \leq 1 - p_0\right] \geq 1 - \alpha$$

A sufficient and smoother condition to satisfy the stochasticity constraint is the following quadratic constraint:

$$\theta^T \sum_s d_\theta([1, s_1, s_2, s_3][1, s_1, s_2, s_3]^T)\theta \leq (\log(\frac{p_0}{1 - p_0}))^2\alpha \tag{3.1}$$

In all of the simulations shown below, we use $\alpha = 0.1$ and $p_0 = 0.1$ unless otherwise specified.

The optimal policy θ^* and the oracle λ^*. Theory established in section 2.1.2 has that for every pair of (p_0, α) there exists a Lagrangian multiplier λ^* such that the optimal

solution to the regularized average cost function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, C(\theta) + \lambda\theta^T \sum_s d_\theta([1, s_1, s_2, s_3][1, s_1, s_2, s_3]^T)\theta \tag{3.2}$$

satisfies the quadratic constraint with equality. Furthermore, λ increases as the stringency of the quadratic constraint: increased value of λ is associated with a decreased value of the quadratic term $\theta^{*T} \sum_s d_{\theta^*}([1, s_1, s_2, s_3][1, s_1, s_2, s_3]^T)\theta^*$. For a fixed pair of (p_0, α), we perform a line search to find the smallest λ, denoted as λ^*, such that the minimizer to the regularized average cost, denoted as θ^* satisfies the quadratic constraint. We recognize the difficulty in solving the optimization problem due to the non-convexity of the regularized average cost function. In search for a global minimizer, we therefore use grid search, for a given λ, to find a crude solution to the optimization problem. We then improve the accuracy of the optimal solution using pattern search function. The regularized average cost function is approximated by Monte Carlo samples. 5000 Monte Carlo samples are used to approximate the regularized average cost for simulation in section 3.1 and 3.2 where the stationary distribution of contexts does not depend on the policy. For the simulations in section 3.3.1 and 3.3.2 where context distribution depends on the policy, we generate a trajectory of 100000 Monte Carlo samples and truncate the first 10% of the samples to approximate the stationary distribution.

Estimating lambda online. In practice the decision maker has no access to the oracle Lagrangian multiplier λ^*. A natural remedy is to integrate the estimation of λ^* with the online actor critic algorithm that estimates the policy parameters. An actor critic algorithm with a fixed Lagrangian multiplier solve the "primal" problem while the "dual" problem searches for λ^*. Our integrated algorithm performs a line search to find the smallest λ such that the estimated optimal policy satisfies the quadratic constraint. The stationary distribution of the contexts is approximated by the empirical distribution. Estimating λ can be very time consuming, therefore in our simulation the algorithm performs the line search on λ every 10 decision points. Similar ideas with gradient based updates on λ have appeared in reinforcement literature to find the optimal policies in constrained MDP problems, see [12, 8] for examples.

Simulation details The simulation results presented in the following sections are based on 1000 independent simulated users. For each simulated user, we allow a burn-in period of 20 decision points. During the burn-in period, actions are chosen by fair coin flips. After the burn-in period, the online actor critic algorithm is implemented to learn the optimal policy and obtain an end-of-study estimated optimal policy at the last decision point. Both bias and MSE shown in all of the following tables are averaged over 1000 end-of-study estimated

optimal policies. For each simulated user the 95% bootstrapped confidence intervals for θ^* is based on 500 bootstrapped samples generated by algorithm 4. We expect with 95% confidence that the empirical coverage rate of a confidence interval should be within 0.936 and 0.964, if the true confidence level is 0.95.

3.1 I.I.D. Contexts

In this generative model, we choose the simplest setting where contexts at different decision points are i.i.d.. We simulate context $\{[S_{t,1}, S_{t,2}, S_{t,3}]\}_{t=1}^{T}$ form a multivariate normal distribution with mean 0 and identity covariance matrix. The population optimal policy is $\theta^* = [0.417778, 0.394811, 0.389474, 0.001068]$ at $\lambda^* = 0.046875$. Table 3.1 and table 3.2 list bias and mean squared error (MSE) of the estimated optimal policy parameters. Both measures shrink towards 0 as T, sample size per simulated user, increases from 200 to 500, which is consistent with the convergence in estimated optimal policy parameter as established in Theorem 2. Table 3.3 shows the empirical coverage rates of percentile-t bootstrap confidence interval at sample sizes 200 and 500. At sample size 200, the empirical coverage rates are between 0.936 and 0.964 for all θ_i's. At sample size 500, however, the bootstrap confidence interval for θ_2 is a little conservative with an empirical coverage rate of 0.968. The symmetric Efron bootstrap confidence intervals are anti-conservative at sample size 200 but have descent coverage at sample size 500, as shown in table 3.4.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	-0.081	-0.090	-0.089	0.010
500	-0.053	-0.037	-0.034	-0.002

Table 3.1: I.I.D. contexts: bias in estimating the optimal policy parameter. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.054	0.052	0.052	0.055
500	0.027	0.024	0.021	0.029

Table 3.2: I.I.D. contexts: MSE in estimating the optimal policy parameter.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.962	0.942	0.938	0.945
500	0.96	0.948	0.968	0.941

Table 3.3: I.I.D. contexts: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.946	0.92*	0.921*	0.939
500	0.947	0.937*	0.952*	0.945

Table 3.4: I.I.D. contexts: coverage rates of Efron-type bootstrap confidence intervals for the optimal policy parameter. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

If we change the estimation goal by imposing a stringent stochasticity constraint, the learning rate of the actor critic algorithm slows down. To see this, we compare two sets of experiments. One is conducted with $\alpha = p_s = 0.1$ in the stochasticity constraint. There is at least 90% of the contexts there is at least 10% chance of selecting both actions. Results of the experiment have been shown in tables 3.1 through table 3.4. The other set of experiment is conducted with $\alpha = 0.2$ and $p_s = 0.1$; that is, under this specification there is at least 10% probability of choosing both actions for at least 80% of the contexts. In a nutshell, we enforce less stochasticity in the optimal policy in the second experiment setting. The Lagrangian multiplier $\lambda^* = 0.046875$ in the first experiment setting while in the second experiment setting we have $\lambda^* = 0.0281$. In the latter setting, minimizing the regularized average cost 3.2 becomes a harder optimization problem due to the lack of curvature of regularized average cost at the optimal policy. A regularized average cost function with a smaller λ^* is "flatter" around the optimal policy. Comparing the curvature of the regularized average cost function at the optimal, Hessian matrix in the first setting has a condition number of 1.25 and determinant $1.9654e-04$ while the Hessian matrix in the second setting has a condition number of 1.30 and a determinant $4.3702e{-}05$. On top of the increased difficulty in optimization, the online actor critic algorithm explores less when the stochasticity constraint is more stringent, which makes the learning of optimal policy less efficient. The combination of these two reasons contribute to a performance degradation of the algorithm in the second experiment. In the second experiment the oracle lambda is $\lambda^* = 0.028$ and the optimal policy is $\theta^* = [0.574245, 0.529603, 0.531282, -0.000]$ which is a more deterministic policy than the optimal policy in the first experiment with

$\theta^* = [0.417778, 0.394811, 0.389474, 0.001068]$ at $\lambda^* = 0.046875$. Table 3.5 and table 3.6 list the bias and MSE in the second experiment setting. Both the bias and MSE diminish towards 0 as sample size increases, albeit at a slower rate than that in the first experiment. Table 3.3 and table 3.7 shows that the percentile-t bootstrap confidence intervals for θ_1 and θ_2 are anti-conservative at sample size 200 while in the first experiment confidence intervals attain descent coverage at the same sample size.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	-0.109	-0.105	-0.117	0.015
500	-0.054	-0.030	-0.038	-0.002

Table 3.5: I.I.D. contexts with a lenient stochasticity constraint: bias in estimating the optimal policy parameter. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.121	0.107	0.104	0.106
500	0.056	0.047	0.046	0.055

Table 3.6: I.I.D. contexts with a lenient stochasticity constraint: MSE in estimating the optimal policy parameter.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.962	0.929*	0.926*	0.95
500	0.968	0.941	0.954	0.951

Table 3.7: I.I.D. contexts with a lenient stochasticity constraint: coverage rates of percentile-t bootstrap confidence interval. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

3.2 AR(1) Context

In this section we study the performance of the actor critic algorithm when the dynamics of the context is an auto-regressive stochastic process. We envision that in many health applications, contexts at adjacent decision points are likely to be correlated. Using HeartSteps as an example, weather (S_1) at two adjacent decisions points are likely to be similar so are users' learning ability (S_2) and disengagement level S_3. One way to incorporate the correlation among contexts at near-by decision points is through a first degree auto-regression process. We simulate the context according to

$$
\begin{aligned}
S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\
S_{t,2} &= 0.4S_{t-1,2} + \xi_{t,2}, \\
S_{t,3} &= \xi_{t,3}
\end{aligned}
$$

Here we choose $\xi_{t,1} \sim N(0, 1 - 0.4^2)$, $\xi_{t,2} \sim N(0, 1 - 0.4^2)$ and $\xi_{t,3} \sim N(0, 1)$ so that the stationary distribution of S_t is multivariate normal with zero mean and identity covariance matrix, same as the distribution of S_t in the previous section. The initial distribution of $S_t, t = 1$ is a multivariate standard normal.

The oracle Lagrangian multiplier is $\lambda^* = 0.05$ and the population optimal policy is $\theta^* = [0.417, 0.395, 0.394, 0]$, same as in the i.i.d. experiment. Bias and MSE of the estimated policy parameters are shown in table 3.8 and table 3.9. Empirical coverage rate of the percentile t bootstrap confidence interval is reported in table 3.10. Both the bias and MSE diminish towards 0 as the sample size increases from 200 to 500, a clear indication that convergence of the algorithm is not affected by the auto-correlation in context. The bootstrap confidence interval for θ_3 is anti-conservative at sample size 200, but recovers descent coverage at sample size 500.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	-0.093	-0.089	-0.076	0.006
500	-0.046	-0.032	-0.040	-0.005

Table 3.8: AR(1) contexts: bias in estimating the optimal policy parameter. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.058	0.053	0.047	0.057
500	0.025	0.022	0.024	0.028

Table 3.9: AR(1) contexts: MSE in estimating the optimal policy parameter

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.963	0.952	0.957	0.927*
500	0.969	0.962	0.96	0.949

Table 3.10: AR(1) contexts: coverage rates of percentile-t bootstrap confidence intervals. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

We continue investigating the influence of auto-correlation on the learning rate, the rate at which MSE in policy estimation shrinks towards 0. To do so, we simulate contexts from the following dynamics and compare the MSE when the auto-regression coefficient η ranges from 0 to 0.9:

$$\begin{aligned} S_{t,1} &= \eta S_{t-1,1} + \xi_{t,1}, \\ S_{t,2} &= \eta S_{t-1,2} + \xi_{t,2}, \\ S_{t,3} &= \xi_{t,3} \end{aligned}$$

Both $S_{t,1}$ and $S_{t,2}$ are generated from first degree auto-regressive process with coefficient η while we leave $S_{t,3}$ as i.i.d.. The noise terms $\xi_{t,1}$ and $\xi_{t,2}$ are independently normally distributed with mean 0 and standard deviation $\sqrt{1-\eta^2}$ so that the long-term stationary distribution of $S_{t,1}$ and $S_{t,2}$ are standard normals. $\xi_{t,3}$ has a standard normal distribution. The auto-regression coefficient η is directly related to the (partial) auto-correlation coefficient and captures the amount of dependency between contexts at adjacent decision points. Figure 3.1 and Figure 3.2 show how relative MSE, relative to the MSE when $\eta = 0$, for the estimated optimal policy parameters changes as the auto-regressive coefficient η increases. The relative MSE for θ_1 and θ_2 has a general increasing pattern as η increases, indicating that stronger auto-correlation among contexts at adjacent decision points slows down the learning rate. MSE for θ_3, the coefficient for S_3, is not grossly affected by the auto-correlation in S_1 and S_2 compared to the other three coefficients.

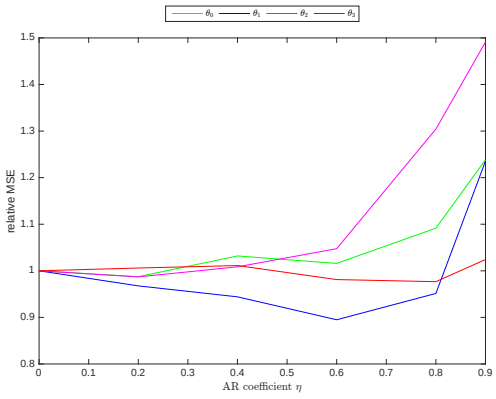

Figure 3.1: Relative MSE vs AR coefficient η at sample size 200. Relative MSE is relative to the MSE at $\eta = 0$.

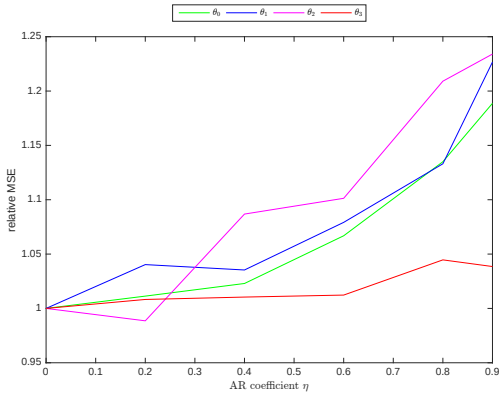

Figure 3.2: Relative MSE vs AR coefficient η at sample size 500. Relative MSE is relative to the MSE at $\eta = 0$.

3.3 Context is Influenced by Previous Actions

We realized that, in many health applications, actions may influence the distribution of contexts, albeit to a minimal extend. For example, the variety of Heartsteps suggestions broaden users' knowledge on how to keep themselves active, which reduces users' sedentary time. We refer to such effect of actions on context as *a learning effect*. On the other hand, if HeartSteps application annoys the users with high volume of activity suggestions, some users may experience *a burden effect*. Burden effects are due to intervention burden. They cause a overall feeling of burden and lack of engagement, which could be eventually reflected on an increase in users' sedentary time. In this section, we first investigate the performance of actor critic algorithm when a learning effect presents. Later we investigate performance of the algorithm when there is a burden effect.

3.3.1 Learning Effect

In this section, we study how actor critic algorithm behaves under a generative model with a learning effect. This generative model represents the type of users who are actively engaged with Heartsteps application and pick up the skills and tactics to stay active as they use the application on a daily basis. To incorporate the learning effect in our generative model, we add a main effect of the previous action in the model of $S_{t,2}$: $S_{t,2}$ increases if there is a physical activity suggestion $A_{t-1} = 1$ at the previous decision point. In this generative model, initial distribution of S_t (t=1) are simulated from a multivariate normal distribution with mean 0 and identity covariance matrix. After the first decision point, contexts are

generated according to the following stochastic process:

$$\begin{aligned} \xi_t &\sim \text{Normal}_4(0, I), \\ S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\ S_{t,2} &= 0.4S_{t-1,2} + \nu A_{t-1} + \xi_{t,2}, \\ S_{t,3} &= \xi_{t,3}, \end{aligned}$$

where ν is a parameter that controls the size of the learning effect. When $\nu = 0$, the context dynamics reduce to a first degree auto-regressive process, the same we investigated in section 3.2. We envision that, in real life, the impact on $S_{t,2}$ from the previous action should not exceed the impact on $S_{t,2}$ from $S_{t-1,2}$, the previous learning ability. Therefore we study the performance of the actor critic algorithm on three types of users with $\nu = 0, 0.2, 0.4$. The three types of users, with $\nu = 0, 0.2, 0.4$, are users with no learning, moderate learning and large learning effect. Table 3.11 lists the optimal policy θ^* and the oracle λ^* at the three different values of ν. Large values of Lagrangian multiplier λ is needed when the size of learning effect increases so that the quadratic constraint is satisfied by the corresponding optimal policy. The optimal policy parameter θ^* follows a pattern that the relative magnitude of θ_0^* compared to the other three coeffients increases as ν increases. This pattern aligns well with our intuition: for the more enthusiastic learner, Heartsteps should recommend physical activity suggestions more often regardless of the context.

ν	λ^*	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.06	0.341	0.327	0.326	0
0.2	0.08	0.481	0.231	0.231	-0.004
0.4	0.11	0.574	0.161	0.165	0

Table 3.11: Learning effect: the optimal policy and the oracle lambda.

Table 3.12, Table 3.13 and Table 3.14 list the bias, mean squared error (MSE) and the empirical coverage rate of the bootstrap confidence interval for the optimal policy parameter θ^* when sample size, the total number of decision points, is 200. Table 3.15, Table 3.16 and Table 3.17 have the same measures for sample size 500. Table 3.12 and Table 3.15 also document the bias in estimating the Lagrangian multiplier online. The bandit actor critic algorithm estimates the optimal policy parameter with low bias and MSE for users with no learning effect ($\nu = 0$). Bootstrap confidence intervals have descent coverage. The results align well with the results obtained from section 3.2. Bias and MSE in estimating

the optimal policy parameter, notably θ_0, θ_1 and θ_2, increase as the learning effect levels up. The bias remains stable as sample increases from 200 to 500. Confidence intervals for θ_0, θ_1 and θ_2 suffer from severe anti-conservatism. Degrading of algorithm is partially due to the fact that bandit actor critic algorithm chooses policies that minimize the average cost and does not take into account the effect of policy on the stationary distribution of contexts. This results in an estimated optimal policy that does not recommend physical activity suggestion as aggressively as one should, which is reflected on an underestimated θ_0 and positive biases in estimating θ_1 and θ_2. In addition to the myopic view of the bandit algorithm, degrading of the algorithm can be partially attributed to the bias in estimating λ online, as shown in table 3.12 and table 3.15. The oracle Lagrangian multiplier λ^* is chosen so that the optimal policy parameter satisfies the quadratic constraint 3.1 while the online bandit actor critic algorithm estimates the Lagrangian multiplier so that the bandit-estimated optimal policy satisfies the quadratic constraint. To separate the consequence of the underestimated λ from the consequence of the myopia of the bandit algorithm, we test the bandit algorithm using a fixed λ^*. Results of those experiments are shown in table 3.35 through table 3.40 in the appendix. There the optimal policy parameters, especially θ_0, are still estimated with large bias and MSE for users with moderate and large learning effect.

Estimation of θ_3, the policy parameter for S_3, is pretty robust to the addition of a learning effect in the generative model. θ_3 is estimated with low bias and MSE, which shrink towards 0 as sample size increases. Moreover, bootstrap confidence intervals for θ_3 have descent coverage rate at different levels of learning effect. Such robustness is critical since in practice it is important to screen out components of the context, such as S_3, that are not useful for personalizing intervention. We conduct additional experiments with correlated $S_{t,2}$ and $S_{t,3}$ by simulating $(\xi_{t,2}, \xi_{t,3})$ from multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = [1, -0.3; \\ -0.3, 1]$$

. We observe that the quality in estimating θ_3 does not change with the introduction of correlation between S_2 and S_3. Results are listed in the appendix from table 3.41 through 3.46.

ν	λ^*	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.00	-0.021	-0.035	-0.034	0.012
0.2	-0.01	-0.163	0.047	0.047	0.016
0.4	-0.04	-0.262	0.104	0.094	0.012

Table 3.12: Learning effect: bias in estimating the optimal policy parameter while estimating λ online at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.048	0.037	0.038	0.044
0.2	0.070	0.035	0.035	0.042
0.4	0.113	0.041	0.037	0.039

Table 3.13: Learning effect: MSE in estimating the optimal policy parameter while estimating λ online at sample size 200.

ν	θ_0	θ_1	θ_2	θ_3
0	0.97	0.958	0.952	0.947
0.2	0.925*	0.934*	0.93*	0.942
0.4	0.864*	0.886*	0.898*	0.941

Table 3.14: Learning effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

ν	λ^*	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.01	0.003	0.011	0.016	-0.002
0.2	-0.02	-0.145	0.086	0.089	0.001
0.4	-0.05	-0.251	0.136	0.134	-0.003

Table 3.15: Learning effect: bias in estimating the optimal policy parameter while estimatingλ online at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.023	0.018	0.016	0.023
0.2	0.042	0.024	0.023	0.020
0.4	0.085	0.034	0.031	0.017

Table 3.16: Learning effect: MSE in estimating the optimal policy parameter while estimating λ online at sample size 500.

ν	θ_0	θ_1	θ_2	θ_3
0	0.98	0.949	0.963	0.954
0.2	0.907*	0.887*	0.889*	0.95
0.4	0.724*	0.777*	0.778*	0.946

Table 3.17: Learning effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. λ is estimated online.Coverage rates significantly lower than 0.95 are marked with asterisks (*).

Another view to the performance of the bandit actor critic algorithm is through the box plot of the regularized average cost of the estimated optimal policy. Figure 3.3 displays three side-by-side box plots, one for each value of ν, of the regularized average cost of the end-of-experiment policies at sample size 200. The three asterisks are the regularized average costs of the optimal policies in table 3.11. Comparing the three types of users, the regularized average cost decreases as the learning effect levels up, an artifact that more learning reduces the sedentary time (cost). The bottom whisker of each box plot stays above the asterisks. The discrepancy between the optimal regularized average cost and the median regularized average costs of the end-of-experiment policies increases as the learning effect elevates, which indicates the worsened quality of the bandit-estimated optimal policy when the size of the learning effect increases. In addition, variance in regularized average costs inflates as the learning effect elevates, a consequence of both increased instability of the algorithm and increased difficulty in solving the optimization problem. Figure 3.4 displays box plots of regularized average costs at sample size 500. As sample size increases, regularized average costs are less variable but their discrepancies from the optimal policy value remain stable.

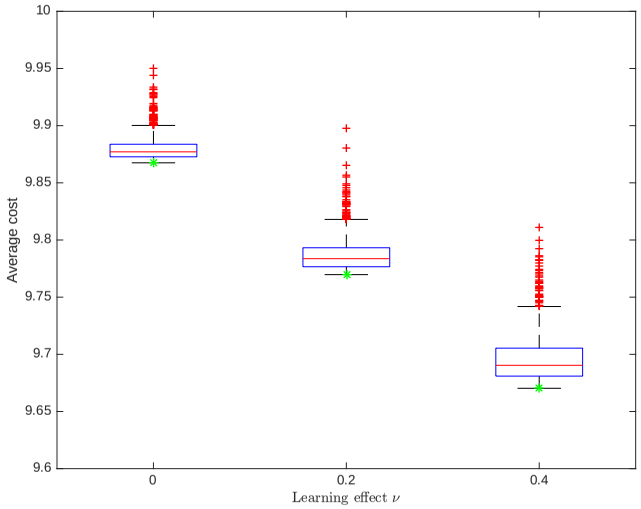

Figure 3.3: Learning effect: box plots of regularized average cost at different levels of learning effect. Sample size is 200.

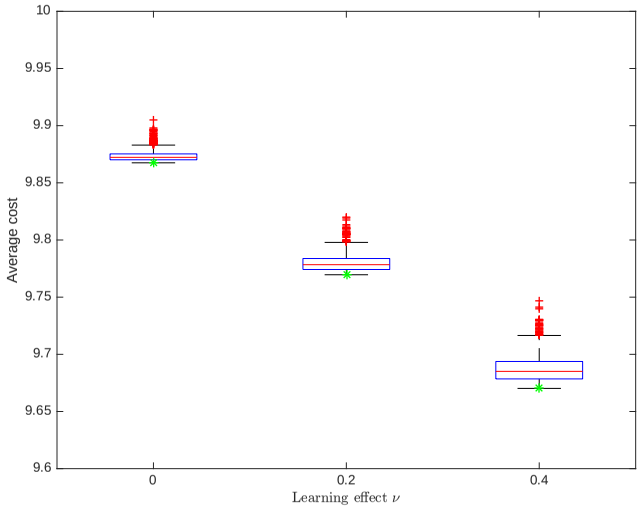

Figure 3.4: Learning effect: box plots of regularized average cost at different levels of learning effect. Sample size is 500.

Although our results show no sign that the optimal policy estimated by the bandit actor critic algorithm will converge to the optimal policy, we observe convergence in the estimated policy as sample size T grows. We conjecture that, when actions affect contexts distributions, the bandit algorithm converges to the policy $\pi_{\theta^{**}}$ that satisfies the following equilibrium equation:

$$\theta^{**} = \underset{\theta}{\operatorname{argmin}} \sum_s d_{\theta^{**}}(s) \sum_a \pi_\theta(A = a|S = s)\mathbb{E}(C|A = a, S = s) - \lambda^{**}\theta^T \mathbb{E}_{\theta^{**}}[g(S)g(S)^T]\theta \tag{3.3}$$

$$\text{where } \lambda^{**} \text{ is the smallest } \lambda \text{ such that } \theta^{**} \sum_s d_{\theta^{**}}(s)g(s)^T g(s)\theta^{**} \leq (\log(\frac{p_0}{1-p_0}))^2\alpha \tag{3.4}$$

When actions do not influence contexts distributions, the equilibrium equation is the same system of equations satisfied by the optimal policy. When previous actions have an impact on context distribution at later decision points, the stationary distribution of context is a function of policy. We call solution to equation 3.4 the myopic equilibrium policy. The myopic equilibrium policy minimizes the regularized average cost under the stationary distribution generated by itself. Such policy achieves an "equilibrium state" and there is no motivation to leave the current policy. The myopic equilibrium policies for different level of learning effect are listed in table 3.18. Table 3.19 through table 3.22 list the bias and MSE in estimating the myopic equilibrium policy when sample size is 200 and 500. Our conjecture is confirmed by results presented in these tables.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.341	0.327	0.326	0
0.2	0.273	0.260	0.260	0
0.4	0.211	0.200	0.200	−0.000

Table 3.18: Learning effect: the myopic equilibrium policy.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.070	-0.079	-0.078	0.013
0.2	-0.054	-0.065	-0.066	0.012
0.4	-0.042	-0.050	-0.057	0.011

Table 3.19: Learning effect: bias in estimating the myopic equilibrium policy at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^{**}$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.052	0.042	0.043	0.044
0.2	0.047	0.037	0.037	0.042
0.4	0.046	0.033	0.032	0.039

Table 3.20: Learning effect: MSE in estimating the myopic equilibrium policy at sample size 200.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.046	-0.032	-0.028	-0.001
0.2	-0.036	-0.025	-0.023	-0.003
0.4	-0.030	-0.018	-0.017	-0.004

Table 3.21: Learning effect: bias in estimating the myopic equilibrium policy at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^{**}$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.025	0.019	0.017	0.023
0.2	0.023	0.017	0.015	0.020
0.4	0.023	0.015	0.014	0.017

Table 3.22: Learning effect: MSE in estimating the myopic equilibrium policy at sample size 500.

3.3.2 Burden Effect

In this section, we study behavior of the actor critic algorithm in the presence of an intervention burden effect. Generative model with a burden effect represents the type of users who disengage with Heartsteps application and the recommend intervention if the application provides physical activity suggestions at high frequency. When users experience

intervention burden effects, they become frustrated and have a tendency of falling back to their unhealthy behavior which leads to an increase in sedentary time. In our burden effect generative model, $S_{t,3}$ represents the disengagement level whose value increases if there is a physical activity suggestion at the previous decision point $A_{t-1} = 1$. The positive main effect of $S_{t,3}$ in the cost model 3.5 reflects that higher disengagement level is associated with higher cost (higher sedentary time). The initial distribution of S_t are simulated from multivariate normal distribution with mean 0 and identity covariance matrix . After the first decision point, contexts are generated according to the following stochastic process:

$$\begin{aligned} S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\ S_{t,2} &= 0.4S_{t-1,2} + \xi_{t,2}, \\ S_{t,3} &= 0.4S_{t-1,3} + 0.2S_{t-1,3}A_{t-1} + 0.4A_{t-1} + \xi_{t,3} \end{aligned}$$

We simulate the cost, sedentary time per hour between two decision points, according to the following linear model:

$$C_t = \quad 10 - .4S_{t,1} - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1} + 0.2S_{t,2}) + \tau S_{t,3} + \xi_{t,0}. \qquad (3.5)$$

where parameter τ controls the "size" of the burden effect: the larger τ is, the more severe burden effect is. We study the performance of on algorithm on five models with $\tau = 0, 0.2, 0.4, 0.6, 0.8$. Different values of τ represent users who experience different levels of burden effect. $\tau = 0$ represents the type of users who experience no burden effect while $\tau = 0.8$ represents the type of users who experience a large burden effect.

Table 3.23 lists the oracle λ^* and the corresponding optimal policy θ^* at different levels of burden effect. Higher level of burden effects calls for increased value of oracle λ^* to keep the desired intervention stochasticity. The negative sign of θ_3^* at $\nu \geq 0.2$ indicates that the application should lower the probability of pushing an activity suggestion when the disengagement level is high. The magnitude of θ_3^* rises with the size of the burden effect, implying that as burden effect increases the application should further lower the probability of pushing activity suggestions at high disengagement level. θ_0^* decreases to be negative when τ increases, which indicates that as the size of burden effect grows, the application should lower the frequency of activity suggestions in general.

ν	λ^*	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.06	0.3410	0.3269	0.3264	0
0.2	0.05	0.0844	0.3844	0.4	-0.1609
0.4	0.06	-0.1922	0.3547	0.3312	-0.2313
0.6	0.08	-0.3312	0.2488	0.2234	-0.2687
0.8	0.1	-0.3883	0.2078	0.2	-0.2687

Table 3.23: Burden effect: the optimal policy and the oracle lambda.

Table 3.24, 3.25 and 3.26 list the bias, MSE and the empirical coverage rate of the percentile-t bootstrap confidence interval at sample size 200. Table 3.27, 3.28 and 3.29 list these three measures at sample size 500. When there is no burden effect ($\tau = 0$), $S_{t,3}$ has no influence on the cost and is therefore considered as a "noise" variable. The optimal policy parameters are estimated with low bias and MSE under the generative model with $\tau = 0$ and the bootstrap confidence intervals have descent coverage, both of which are clear indications that the algorithm is robust to presence of noise variables that are affected by previous actions. As burden effects level up, we observe an increased bias and MSE in the estimated optimal policy parameters, θ_0 and θ_3 in particular. The empirical coverage rates of bootstrap confidence intervals for θ_0 and θ_3 are below the nominal 95% level. There are two reasons to explain the increased bias and MSE. The most important one is the near-sightedness of bandit actor critic algorithm. The bandit algorithm chooses the policy that maximizes the (immediate) average cost while ignoring the negative consequence of a physical activity suggestion $A_t = 1$ on the disengagement level at the next decision point. The bandit algorithm therefore tends to "over-treat" in general and in particular at high disengagement level, which is reflected in an over-estimated θ_0 and θ_3. The second reason comes from the bias in estimating λ, the Lagrangian multiplier. The oracle Lagrangian multiplier λ^* is chosen so that the optimal policy parameter satisfies the quadratic constraint 3.1 while the online bandit actor critic algorithm estimates the Lagrangian multiplier so that the bandit-estimated optimal policy satisfies the quadratic constraint. To separate the consequence of underestimated λ from the consequence of the myopia of the bandit algorithm, we implement the bandit algorithm with oracle λ^*. Results of these experiments are shown in table 3.47 through table 3.52 in the appendix. We observe that, even with the use of oracle λ^*, the overestimation of θ_0 and θ_3 as well as the anti-conservatism of the confidence intervals are still present.

Overall, the estimation of θ_1 and θ_2 shows robustness to the presence of burden effects. θ_1 and θ_2 are estimated with low bias and MSE under the presence of small to moderate burden effects ($\tau = 0.2, 0.4$). While we observe biases in estimating θ_1 and θ_2 under

moderate to large burden effects ($\tau = 0.6, 0.8$), the magnitude of such bias increases slowly with the size of the burden effect. Empirical coverage rates of the bootstrap confidence intervals for θ_1 and θ_2 are descent for $\tau = 0.2, 0.4$ and only degrades slowly under 95% when $\tau = 0.6, 0.8$.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.027	-0.036	-0.030	0.003
0.2	0.229	-0.093	-0.104	0.164
0.4	0.506	-0.063	-0.035	0.235
0.6	0.645	0.043	0.073	0.272
0.8	0.702	0.084	0.096	0.272

Table 3.24: Burden effect: bias in estimating the optimal policy parameter while estimating λ online at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.058	0.037	0.036	0.036
0.2	0.110	0.044	0.046	0.063
0.4	0.313	0.040	0.037	0.091
0.6	0.473	0.038	0.041	0.110
0.8	0.550	0.043	0.045	0.110

Table 3.25: Burden effect: MSE in estimating the optimal policy parameter while estimating λ online at sample size 200.

τ	θ_0	θ_1	θ_2	θ_3
0	0.963	0.963	0.955	0.942
0.2	0.853*	0.946	0.937	0.862*
0.4	0.565*	0.96	0.954	0.776*
0.6	0.39*	0.937	0.916*	0.739*
0.8	0.329*	0.908*	0.899*	0.739*

Table 3.26: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.006	0.010	0.017	-0.008
0.2	0.263	-0.048	-0.057	0.153
0.4	0.539	-0.018	0.012	0.224
0.6	0.678	0.088	0.120	0.261
0.8	0.735	0.129	0.143	0.261

Table 3.27: Burden effect: bias in estimating the optimal policy parameter while estimatingλ online at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.027	0.018	0.016	0.019
0.2	0.096	0.020	0.019	0.042
0.4	0.318	0.018	0.016	0.069
0.6	0.487	0.026	0.030	0.087
0.8	0.568	0.035	0.037	0.087

Table 3.28: Burden effect: MSE in estimating the optimal policy parameter while estimating λ online at sample size 500.

τ	θ_0	θ_1	θ_2	θ_3
0	0.973	0.949	0.955	0.942
0.2	0.714*	0.95	0.962	0.788*
0.4	0.217*	0.951	0.961	0.635*
0.6	0.101*	0.886*	0.835*	0.545*
0.8	0.07*	0.806*	0.788*	0.546*

Table 3.29: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

Figure 3.5 and 3.6 assess the quality of the estimated optimal policies by comparing the regularized average cost with the optimal regularized average cost in table 3.23. Figure 3.5 does the comparison at five levels of burden effect: $\tau = 0, 0.2, 0.4, 0.6, 0.8$, at sample size 200. As the burden effects level up, the overall long-run average cost goes up, which is simply an artifact of the increasing main effect size of the disengagement level. Having a higher long-term average cost, the estimated optimal policy by the contextual bandit

algorithm is always inferior then the optimal policy. The gap of inferiority, as measure by the difference between the median long-run average cost and the long-run average cost of the optimal policy increases as τ increases. When sample size increases from 200 to 500, we observe less variation in the long-run average cost of the estimated optimal policies. Nevertheless, the gap of inferiority remains stable .

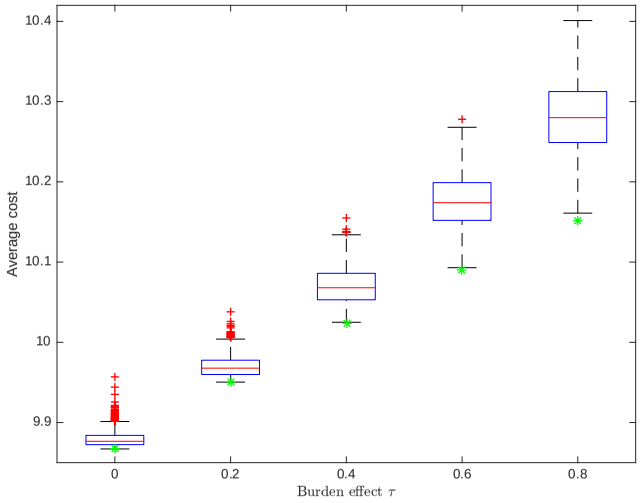

Figure 3.5: Burden effect: box plots of regularized average cost at different levels of the burden effect at sample size 200.

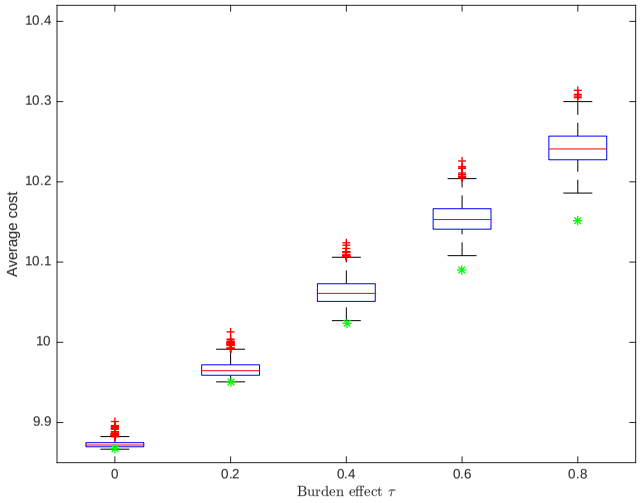

Figure 3.6: Burden effect: box plots of regularized average cost at different levels of the burden effect at sample size 500.

The conjecture regarding the convergence of our bandit algorithm in a full-blown MDP is again supported by results shown in table 3.30 through table 3.34. Table 3.30 lists the solution to the myopic equilibrium system of equations 3.4. Solution remains unchanged at different levels of burden effect since the underlying contexts dynamics is unchanged at different levels of burden effect. The shrinking bias (table 3.31 and table 3.33) and MSE (table 3.32 and 3.34) are consistent with our conjecture that the estimated optimal policy by the bandit algorithm converges to the myopic equilibrium policy.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.392	0.372	0.371	-0.001
0.2	0.392	0.372	0.371	-0.001
0.4	0.392	0.372	0.371	-0.001
0.6	0.392	0.372	0.371	-0.001
0.8	0.392	0.372	0.371	-0.001

Table 3.30: Burden effect: the myopic equilibrium policy.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.078	-0.081	-0.075	0.004
0.2	-0.078	-0.081	-0.075	0.004
0.4	-0.078	-0.081	-0.075	0.004
0.6	-0.078	-0.081	-0.075	0.004
0.8	-0.078	-0.081	-0.075	0.004

Table 3.31: Burden effect: bias in estimating the myopic equilibrium policy at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^{**}$

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.063	0.042	0.041	0.036
0.2	0.063	0.042	0.041	0.036
0.4	0.063	0.042	0.041	0.036
0.6	0.063	0.042	0.041	0.036
0.8	0.063	0.042	0.041	0.036

Table 3.32: Burden effect: MSE in estimating the myopic equilibrium policy at sample size 200.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.045	-0.036	-0.028	-0.007
0.2	-0.045	-0.036	-0.028	-0.007
0.4	-0.045	-0.035	-0.028	-0.007
0.6	-0.045	-0.035	-0.028	-0.007
0.8	-0.045	-0.036	-0.028	-0.007

Table 3.33: Burden effect: bias in estimating the myopic equilibrium policy at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^{**}$

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.029	0.019	0.017	0.019
0.2	0.029	0.019	0.017	0.019
0.4	0.029	0.019	0.017	0.019
0.6	0.029	0.019	0.017	0.019
0.8	0.029	0.019	0.017	0.019

Table 3.34: Burden effect: MSE in estimating the myopic equilibrium policy at sample size 500.

3.4 Appendix

3.4.1 Learning Effect: Actor Critic Algorithm Uses λ^*

The following tables present, when there is a learning effect, the simulation results from running the actor critic algorithm that uses λ^* throughout.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	-0.004	-0.025	-0.025	0.011
0.2	-0.202	0.012	0.010	0.012
0.4	-0.355	0.029	0.021	0.005

Table 3.35: Learning effect: bias in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.061	0.039	0.040	0.042
0.2	0.082	0.026	0.026	0.028
0.4	0.152	0.017	0.016	0.017

Table 3.36: Learning effect: MSE in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0.0	0.948	0.939	0.935*	0.947
0.2	0.856*	0.936	0.929*	0.945
0.4	0.433*	0.94	0.926*	0.944

Table 3.37: Learning effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	-0.019	-0.014	-0.008	-0.003
0.2	-0.219	0.018	0.023	0.001
0.4	-0.373	0.032	0.032	-0.003

Table 3.38: Learning effect: bias in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.025	0.017	0.016	0.019
0.2	0.064	0.011	0.011	0.012
0.4	0.148	0.008	0.007	0.007

Table 3.39: Learning effect: MSE in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0.0	0.956	0.940	0.949	0.955
0.2	0.613*	0.932*	0.932*	0.946
0.4	0.035*	0.916*	0.913*	0.945

Table 3.40: Learning effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

3.4.2 Learning Effect with Correlated S_2 and S_3: Actor Critic Algorithm Uses λ^*

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.020	-0.034	-0.034	0.011
0.2	-0.160	0.048	0.049	0.016
0.4	-0.262	0.106	0.096	0.009

Table 3.41: Learning effect with correlated S_2 and S_3: bias in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.048	0.037	0.043	0.048
0.2	0.070	0.036	0.040	0.046
0.4	0.115	0.042	0.041	0.041

Table 3.42: Learning effect with correlated S_2 and S_3: MSE in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0	0.972	0.963	0.95	0.952
0.2	0.926*	0.934*	0.928*	0.944
0.4	0.859*	0.893*	0.892*	0.941

Table 3.43: Learning effect with correlated S_2 and S_3: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

ν	θ_0	θ_1	θ_2	θ_3
0	0.005048	0.011879	0.014148	-0.001817
0.2	-0.14299	0.089469	0.08766	0.002411
0.4	-0.25172	0.13707	0.13359	-0.003746

Table 3.44: Learning effect with correlated S_2 and S_3: bias in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$

ν	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.023	0.018	0.018	0.025
0.2	0.042	0.025	0.024	0.022
0.4	0.085	0.033	0.033	0.019

Table 3.45: Learning effect with correlated S_2 and S_3: MSE in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0	0.983	0.95	0.967	0.952
0.2	0.895*	0.876*	0.903*	0.953
0.4	0.717*	0.773*	0.791*	0.949

Table 3.46: Learning effect with correlated S_2 and S_3: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

3.4.3 Burden Effect: Actor Critic Algorithm Uses λ^*

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	-0.027	-0.036	-0.030	0.003
0.2	0.229	-0.093	-0.104	0.164
0.4	0.506	-0.063	-0.035	0.235
0.6	0.645	0.043	0.073	0.272
0.8	0.702	0.084	0.096	0.272

Table 3.47: Burden effect: bias in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.058	0.037	0.036	0.036
0.2	0.110	0.044	0.046	0.063
0.4	0.313	0.040	0.037	0.091
0.6	0.473	0.038	0.041	0.110
0.8	0.550	0.043	0.045	0.110

Table 3.48: Burden effect: MSE in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0	0.963	0.963	0.955	0.942
0.2	0.853*	0.946	0.937	0.862*
0.4	0.565*	0.96	0.954	0.776*
0.6	0.39*	0.937	0.916*	0.739*
0.8	0.329*	0.908*	0.899*	0.739*

Table 3.49: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	-0.018	-0.014	-0.006	-0.009
0.2	0.288	-0.031	-0.040	0.149
0.4	0.516	-0.042	-0.011	0.223
0.6	0.591	0.005	0.037	0.262
0.8	0.606	0.006	0.020	0.263

Table 3.50: Burden effect: bias in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.029	0.017	0.015	0.016
0.2	0.121	0.022	0.021	0.042
0.4	0.294	0.018	0.016	0.066
0.6	0.367	0.011	0.012	0.079
0.8	0.379	0.008	0.008	0.076

Table 3.51: Burden effect: MSE in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0.0	0.944	0.950	0.952	0.933*
0.2	0.689*	0.943	0.959	0.815*
0.4	0.159*	0.944	0.954	0.6*
0.6	0.006*	0.941	0.928*	0.295*
0.8	0*	0.94	0.944	0.144*

Table 3.52: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

CHAPTER 4

A Multiple Decision Procedure for Personalizing Intervention

Increasing pharmaceutical and medical research are focusing on developing personalized intervention/medicine that is targeted to a specific subgroup of patients. There is substantial evidence on the heterogeneity in molecular pathogenesis and intervention responses. Personalized intervention utilizes a decision rule that inputs patients' characteristics and outputs a prescription given a set of candidate interventions. Clinical trials which recruit highly heterogenous patients usually record a large amount of baseline patient information, which can be useful inputs in personalizing treatment. Collecting such information, however, may be expensive or time-consuming in real clinical settings. Therefore statistical methodology needs to be developed to identify information useful for personalizing intervention. Many statisticians have contributed works on developing statistical methodology for personalized treatment with the goal of extracting (a combination of) useful variables from a (high-dimensional) set of baseline variables (for example, [30], [14]).

The goal of this chapter is to develop hypothesis testing method for personalizing treatments. We focus on identifying the usefulness of a particular patient characteristic, referred to as biomarker in the following discussion. We define a discrete-valued biomarker as useful in personalizing treatment if for a particular value of the biomarker, there is sufficient evidence to recommend one treatment, while for other values of the biomarker, either there is sufficient evidence to recommend a different treatment, or there is insufficient evidence to recommend a particular treatment. This definition generalizes the concept of qualitative interaction in [28], where a biomarker is deemed useful only if there is sufficient evidence that the recommended treatments varies given different values of the biomarker. It is worth pointing out that [71], [73] also recognized that qualitative interaction is not the only type of interaction useful for personalizing decision making. They redefined qualitative interaction by saying that "a qualitative interaction does require a reversal of effect, but includes situations where there is a treatment effect for one subset and no treatment effect for an-

other". In my following discussion, I will to the definition of qualitative interaction in [28] as *restricted qualitative interaction*.

We consider the scenario where the biomarker is binary and thus divides the patients into subgroup 1 and subgroup 2. We also assume that there are two candidate treatments, treatment A and treatment B. The mean treatment response in subgroup i under treatment X is denoted by μ_{iX}, where $i \in \{1, 2\}$, $X \in \{A, B\}$. The treatment effects are denoted by $\theta_1 = \mu_{1A} - \mu_{1B}$ and $\theta_2 = \mu_{2A} - \mu_{2B}$, respectively in subgroup 1 and subgroup 2. The null hypothesis that the biomarker is not useful for personalizing treatment is

$$H : \theta_1 = \theta_2 = 0, \text{ or } \theta_1\theta_2 > 0 \tag{4.1}$$

Given $\theta_1 = \theta_2 = 0$, there is not enough evidence to demonstrate a treatment effect in neither subgroup. Given $\theta_1\theta_2 > 0$, the same treatment should be recommended regardless of the value of the biomarker. Therefore the biomarker is useful in personalized decision making in neither of the two scenarios.

The alternative hypothesis that the biomarker is useful for personalizing treatment is the complement of H:

$$K : \theta_1 = 0, \theta_2 \neq 0, \text{ or } \theta_1 \neq 0, \theta_2 = 0, \text{ or } \theta_1\theta_2 < 0 \tag{4.2}$$

Under the scenario where $\theta_1 = 0, \theta_2 \neq 0$ or $\theta_1 \neq 0, \theta_2 = 0$, a particular treatment is recommended to one subgroup of patients while the other subgroup, factors such as local considerations, such as costs, side effects and preferences can be the deciding factor in choosing a treatment. We call this scenario *a generalized qualitative interaction*. Given $\theta_1\theta_2 < 0$, the existence of *a restricted qualitative interaction*, different treatments should be enforced in different subgroups.

The following illustrative example for personalizing treatment in treating ADHD children is based on Adaptive Pharmacological and Behavioral Treatments for Children with ADHD Trial (Pelham, personal communication). A potential biomarker is children's history of ADHD medication use. Assign biomarker value 1 to medication naive children and assign value 2 to children with a previous ADHD medication intake. The two active treatments are medication and behavioral intervention. The existence of a qualitative interaction ($\theta_1\theta_2 < 0$) suggests that different treatment ought to be prescribed based children's prior medication use. Suppose that, however, medication and behavioral intervention may not appear to work differently for ADHD children with a prior medication use, whereas for medication naive children, behavioral intervention has a positive treatment effect over medication. Knowing that a child has previously taken medication, in this case, provides

the decision makers to the freedom to choose based on local considerations. For instance, parents who object to medications due to the side-effects or the potential increasing dose down the line may choose behavioral intervention while parents who are less reluctant to utilize time-consuming treatments may opt for medication.

This chapter is organized as the following. In section 4.1, we provide a brief literature review on the related work including the test of restricted qualitative interaction, multiple hypothesis testing and multiple decision theory. Our philosophy and our methodology is motived by some of the existing works. In section 4.2, we propose a two-stage testing procedure for the hypothesis testing problem with null hypothesis H and alternative hypothesis K. In the end of the chapter, we discuss generalization of the current methods and future works.

4.1 Literature Review

4.1.1 The test of qualitative interaction

The null hypothesis that [28] used for detecting qualitative interaction is $H_G : \theta_1\theta_2 \geq 0$. Assuming normality and known variances, they developed a likelihood ratio test for the hypothesis. The maximum likelihood estimators of θ_1 and θ_2 are denoted by $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. The test statistic takes the form:

$$T_{GS} = \min\{\max\{\hat{\theta}_1, \hat{\theta}_2\}, \max\{-\hat{\theta}_1, -\hat{\theta}_2\}\} \tag{4.3}$$

The critical value is chosen to control the type I error rate at the least favorable configurations (LFC), which are $(\theta_1, \theta_2) = (0, \infty), (0, -\infty), (\infty, 0), (-\infty, 0)$. They've shown that, when there are two subgroups, the critical value for size α test is z_α, the upper 100α percentile of the standard normal distribution.

One of the main criticisms Gail and Simon's likelihood ratio test has received is its poor power. The test is biased, both in finite sample and asymptotically, in the sense that the power function evaluated at the alternative space may be lower than the size of the test. Asymptotically, the power of Gail and Simon's test is close to $2\alpha^2$ in places near the origin in the alternative space. Bias of the likelihood ratio test indicates that no matter how large the sample size is, there will always exist points in the alternative space at which the probability of correctly rejecting the null hypothesis may be smaller than the probability of a false rejection. The bias of test only gets worse when the number of subgroups increases: the power of the test near the origin decreases exponentially when the number of subgroups

increases.

A few authors have published work in an attempt to improve the power of Gail and Simon's test. [63] proposed a range test for detecting the qualitative interactions. The range and the likelihood ratio test are identical when there are two subgroups. The range test outperforms Gail and Simon's likelihood ratio test, if the number of subgroups is more than two and the signs of the treatment effects are consistent in the majority of subgroups (for example, 80% of the subgroups). In all other scenarios the likelihood ratio tests has better power. [5] and [93] proposed hypothesis testing procedures which can be applied to the testing of qualitative interactions when there are two subgroups. The power of their new methods dominate that of Gail and Simon's. Both methods carefully enlarge Gail and Simon' rejection region while controlling for the type I error rate. Both methods, however, received criticism from [62], who argued that both methods are counter-intuitive, for the rejection regions are not monotone and include samples that are arbitrarily close to the null space.

[32] summarized the challenges in hypothesis testing in which the null is a composite hypotheses about a vector of parameters. The lack of pivotal quantities and possibly the dependency of the distributions of test statistics on nuisance parameters motivated the use of least favorable configuration. The such-derived critical value, which is based on the distribution of the test statistic at the LFC, is a conservative. The power of such tests are inevitably sacrificed at parameter values far away from the LFC. Hansen proposed a testing method with improved power based on data-driven LFC. In stead of searching the entire null space for the LFC, he used the data to narrow down the search. Let θ be the parameter of interest and $\theta \in \Theta_0$ be the null hypothesis. The old way to calculate the critical value, given a test statistic T, is to take the supreme of all upper α percentile of the distribution of T, with the supremum being taken over the entire Θ_0. Hansen proposed to first estimate θ by $\hat{\theta}_n$ and define $C_\epsilon \equiv \mathcal{N}_\epsilon(\hat{\theta}_n) \cap \Theta_0$, where $N_\epsilon(\ )$ is the ϵ neighborhood and n is the sample size. The data-dependent critical value is the supremum of all upper α percentile of the distribution of T, with supremum taken over C_ϵ. The power of Hansen's test procedure dominates that of the LFC test. He provided guidance on how to choose ϵ as a function of n and proved that the test is asymptotically similar on the boundary $\partial\Theta_0$. The idea of data-driven critical values has gained attention in both statistical and econometrical societies. See [6], [51]) for examples.

4.1.2 Multiple Hypothesis Testing, Multiple Decision Theory

The following two articles [47], [48] set the foundation for multiple decision theory and bridged multiple decision problem with hypothesis testing problem. In the very beginning of the first article, Lehmann compared the pros and cons of formulating a statistical inference procedure as a hypothesis testing problem and a multiple decision problem:

"*One of the attractions of formulating statistical problems in terns of hypothesis testing is the resulting structural simplicity. However, at the same time this reduction to a choice between only two decisions frequently causes complications by creating a class of alternatives which combines too many different elements. In many such cases, if one is willing to forego structural simplicity and to divide the class of alternatives into its- natural components, one obtains a multiple decision problem, which admits a simpler and more natural solution than the apparently less complex testing problem.*"

We resonate with Lehmann's message. Often, the alternative hypothesis is comprised of different components, each of which may lead to a remarkably different consequence. By using an accept-reject decision rule one implicitly treats different components in the alternative as if they impact the real-life problem in similar way.s This oversimplification can be misleading and can cause difficulty in interpreting the decisions. In our testing problem, the alternative space consists of three parts $\{\theta : \theta_2 = 0, \theta_1 \neq 0\} \bigcup \{\theta : \theta_1 \neq 0, \theta_2 = 0\} \bigcup \{\theta : \theta_1\theta_2 < 0\}$. When the null hypothesis H is rejected, it is desirable to make finer conclusion on which part of the alternative space θ belongs to, since in the three different scenarios we form different decision rules in recommending personalized treatment. If the conclusion is $\{\theta : \theta_2 = 0, \theta_1 \neq 0\}$, we recommend to conduct follow-up study to confirm the treatment effect in subgroup 1 while the recommendation in subgroup 2 can be based on local considerations. On the other side, if the conclusion is $\{\theta : \theta_1\theta_2 < 0\}$, two more clinical trials should be conducted to confirm the crossover treatment effects in the two subgroups. By forming the problem as a multiple decision problem, we are able to make finer decisions than an oversimplified accept-reject decision.

Lehmann considered a multiple decision problem induced by simultaneously testing a family of hypotheses $\{H_\gamma : \theta \in \omega_\gamma\}$ where $\gamma \in \Gamma$. Different decisions corresponds to different statements regarding which of the hypotheses are false and which of them are true. The family of hypotheses partitions the parameter space into what Lehmann called "atoms". Each atom is defined by

$$\Omega_i = \bigcap_{\gamma \in \Gamma} \omega_\gamma^{x_{i\gamma}}$$

where $x_{i_\gamma} = 1$ if hypothesis H_γ is true and $x_{i_\gamma} = -1$ if hypothesis H_γ is false for the given Ω_i. The loss function, when the true $\theta \in \Omega_i$ but the decision is that $\theta \in \Omega_k$, is

$$\omega_{ik} = \sum_{\gamma \in \Gamma} (\epsilon_{ik\gamma} a_\gamma + \epsilon_{ki\gamma} b_\gamma)$$

where $\epsilon_{ik\gamma}$ equals 1 if $x_{i\gamma} = 1$, $x_{k\gamma} = -1$ and 0 otherwise. The loss function is additive in the sense that it sums up all the losses for making Type I errors and Type II errors.

Lehmann proved a main theorem in [47]. Suppose that for each hypothesis H_γ, the test φ_γ^0 uniformly minimizes the risk among all tests that are similar on the boundary at level $\alpha_\gamma = \frac{b_\gamma}{a_\gamma + b_\gamma}$ and that the family $\{\varphi_\gamma^0, \gamma \in \Gamma\}$ is compatible. Under certain regularity conditions, the product procedure is unbiased and uniformly minimizes the risk among all unbiased decision procedure of the product problem, assuming the same loss function is used.

4.1.2.1 The three principles

There are three fundamental principles the multiple hypotheses and multiple decision theory: the closure principle, the partitioning principle and the sequential rejection principle. Before explaining the principles, let's first recall the (strong) control of the family wise error rate (FWER).

The strong control of FWER ([49]): Given a family of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and a family of hypotheses indexed by I: $\mathcal{H} = \{H_i\}_{i \in I}$, a multiple test $\psi = \{\psi_i\}_{i \in I}$ is said to control the FWER in the strong sense if

$$\forall \theta \in \Theta : P_\theta(\bigcup_{i \in I(\theta)} \{\psi_i = 1\}) \leq \alpha \tag{4.4}$$

where $I(\theta) = \{i \in I : \theta \in H_i\}$. An equivalent definition is that

$$\forall \emptyset \neq J \subset I : \forall \theta \in H_J = \bigcap_{j \in J} H_j : P_\theta(\bigcup_{j \in J} \{\psi_j = 1\}) \leq \alpha \tag{4.5}$$

In contrast, ψ is said to control the FWER in the weak sense if

$$\forall \theta \in H_I = \bigcap_{i \in I} H_i : P_\theta(\bigcup_{i \in I} \{\psi_i = 1\}) \leq \alpha \tag{4.6}$$

The closure principle first appeared in [53] who considered simultaneous testing of a family of hypotheses $\mathcal{H} = \{H_i\}_{i \in I}$ that is closed under intersection. Specifically, the

intersection $\bigcap_{i\in J} H_i$ is either empty or belongs to $\mathcal{H}$ for any $J \subset I$. A natural requirement for a decision rule is **coherence**, meaning that if H_i is accepted and $H_i \subset H_j$, H_j should also be accepted. The closure principle says that, if a decision rule $\psi = \{\psi\}_{i\in I}$ is coherent and each ψ_i controls the type I error rate for component hypothesis H_i, ψ controls the familywise error rate for testing $\mathcal{H}$ in the strong sense. One way to interpret the short proof given in [53] is that the coherence property guarantees that controlling the FWER for $\mathcal{H}$ amounts to controlling the type I error rate for the **global** hypothesis $\bigcap_{i\in I} H_i$. Another way to put it is that the strong and the weak control of FWER is equivalent for a coherent test. A theorem which is originally presented in [74] showed that given any multiple test ψ which strongly controls the FWER at α can be "conherentized" by defining $\bar{\psi}$ with $\bar{\psi}_i = \max_{j:H_j\supseteq H_i} \psi_j$. The resulting test still controls the FWER in the strong sense and is at least as large as ψ which may lead to better power. Last but not least, when the hypotheses of interest $\mathcal{H}$ is not closed under intersection, one can consider the smallest closure of $\mathcal{H}$ without additional cost due to the one-to-one correspondence between a coherent multiple level α test for $\mathcal{H}$ and that for the "closure" of $\mathcal{H}$.

In general, when the intersection of a subset of hypotheses is empty, the closure principle still applies. Starting from the set of hypotheses of primary interest $\mathcal{H}$, one generate its closure $\bar{\mathcal{H}}$ which contains all non-empty intersection of a subset of hypotheses from $\mathcal{H}$. Given $H_i, H_j \in \bar{\mathcal{H}}$, we call H_j a *descendant hypothesis* of H_i if $H_j \subset H_i$, which means that H_j is generated by intersecting H_i with some other hypotheses. H_i is an *ascendant hypothesis* of H_j. When a hypothesis does not have any descendants, it's *minimal*. For example, the global hypothesis in the last paragraph is a minimal hypothesis. Notice that all minimal hypotheses are disjoint and that a multiple test of a disjoint family of hypotheses controls the FWER at level α if and only if it controls the type I error rate at level α for each component hypothesis. The closed testing procedure begins by testing all minimal hypotheses at level α. One proceeds to test a hypothesis H_i if all of its descendent hypotheses are rejected; otherwise H_i is automatically accepted.

The partitioning principle was proposed by [27] upon noticing that for a family of disjoint hypotheses, a test ψ has multiple level α if and only if every component ψ_i has level α for testing H_i. Naturally, if one partition the union of all hypotheses $\bigcup_{i\in I} H_i$ into a set of disjoint base hypotheses $\{\Theta_i\}$ so that each H_i can be written as the sum of some base hypotheses. Finding the level α test for each H_i and then "coherentizing" (by applying the closure principle) can now be replaced by finding level α tests for each Θ_i followed by "coherentizing". A natural partition (the coarsest partition) for a closed family $\mathcal{H}$ is given by $\Theta(J_p) = \{\Theta_i : i \in J_p\}$, where $\Theta_i = H_i \bigcap (\bigcup_{j:H_j\subset H_i} H_j)^c$ and $J_p = \{i \in I : \Theta_i \neq \emptyset\}$.

In general, when determining a rejection rule which controls the type I error rate one usually looks for the LFC of the test statistic over the null hypotheses. Restricting the LFC to Θ_i as opposed to H_i may lead to a less conservative rejection rule and possibly an increased power.

[29] proposed **the sequential rejection principle** of familywise error control. The general sequential rejective multiple testing procedure encompasses many well-known methods including those based on the closure principle and the partitioning principle. One important feature of sequential procedures is that decision of rejection made at one step depends on the set of hypotheses rejected in the previous steps. Rejection of hypotheses make the rejections of the remaining easier. Another notable feather is that at each step it is only necessary to control the FWER with respect to the distributions under which all previous rejections are correct rejections (i.e., assuming no type I error has been made). Specifically, in testing a family of hypotheses $\mathcal{H}$, any sequential procedure can be described by a random and measurable function $\mathcal{N}$ which maps the power set $2^{\mathcal{H}}$ to itself. At each step, this function inputs the set of rejected hypotheses and outputs what to reject next. Let $\mathcal{R}_i \subseteq \mathcal{H}$ be the set of hypotheses rejected up till step i, and

$$\mathcal{R}_0 = \emptyset \tag{4.7}$$

$$\mathcal{R}_{i+1} = \mathcal{R}_i \bigcup \mathcal{N}(\mathcal{R}_i) \tag{4.8}$$

Let $\mathcal{R}_\infty = \bigcup_i \mathcal{R}_i$ be the final collection of rejected hypotheses. Goeman and Solari proved that, under *monotonicity condition* that for every $\mathcal{R} \subseteq \mathcal{S} \subset \mathcal{H}$, almost surely

$$\mathcal{N}(\mathcal{R}) \subseteq \mathcal{N}(\mathcal{S}) \bigcup S \tag{4.9}$$

and the *single step condition* that for every θ

$$P_\theta(\mathcal{N}(\mathcal{F}(\theta)) \subseteq \mathcal{F}(\theta)) \geq 1 - \alpha \tag{4.10}$$

Then for every θ,

$$P_\theta(\mathcal{R}_\infty \subseteq \mathcal{F}(\theta)) \geq 1 - \alpha \tag{4.11}$$

In the above, $\theta \in \Theta$ is the parameter of interest and $\mathcal{F}(\theta)$ and $\mathcal{T}(\theta)$ is the set of false hypotheses and the set of true hypotheses when the underlying value of the parameter is θ.

4.2 The Decision Procedure and Controlling the Error Probabilities

Our aim is to develop a statistical decision making procedure for personalizing treatment based on data collected from two arm randomized clinical trials. We achieve the following objectives:

- When the null hypothesis that the biomarker is not useful for personalizing treatment is rejected, the decision procedure distinguishes whether or not there is a sufficient evidence to demonstrate a qualitative interaction, if not, identifies in which subgroup there are evidence to demonstrate a treatment effect. In other words, the decision procedure identifies the type of qualitative interaction.
- The power of detecting a restricted qualitative interaction of the proposed procedure, is at least as large as Gail and Simon's likelihood ratio test.

4.2.1 Notation and Assumptions

Our data comes from two-arm randomized clinical trials which compare two active treatments, treatment A and treatment B. The biomarker is measured as a baseline variable taking value in $\{1, 2\}$. Patients with biomarker value i consists of subgroup i, for $i \in \{1, 2\}$. We use p_1 and p_2 to denote the fractions of the two subgroups in the overall population. We use n to denote the total sample size and n_i to denote the sample size in subgroup i. Subgroup treatment effects are denoted by θ_1 and θ_2 as aforementioned. We assume that p_1 and p_2 are known for the moment. We also assume that, for simplicity, the proportion of patients who are randomized to treatment A is $\frac{1}{2}$ in each subgroup. This can be approximately guaranteed by block randomization. Last but not least, we assume that the sample fraction of patients in subgroup i, $\frac{n_i}{n}$, is equal to the population fraction p_i.

We assume that the distribution of the treatment responses, among subgroup i who are assigned treatment T, $T \in \{A, B\}$, follow a normal distribution with mean μ_{iT} and unit variance.

4.2.2 The Decision Space

Compared to the standard statistical hypothesis testing problem in which one either accepts or rejects the null hypothesis, our procedure follows the paradigm in [47] and [48] and recognizes that different parameters in the alternative space K lead to different clinical decisions. For example, when the true parameter belongs to $\{\theta : \theta_1 = 0, \theta_2 \neq 0\}$, the clinical

implication is to suggest follow-up study in subgroup 2 to verify the treatment effect. On the other hand, when the true parameter belongs to $\{\theta : \theta_1\theta_2 < 0\}$, the clinical implication might be to suggest follow-up studies in both subgroups to confirm the detected treatment effects and the qualitative interaction. The decision space, denoted by $\mathcal{D}$, contains the four decisions that are summarized in Table 4.1. Clinical decision 1 corresponds to accept the null hypothesis H and conclude that there is not sufficient evidence that the biomarker is useful for personalizing decision making, while decision 2, 3, and 4 correspond to reject H and conclude that the biomarker may be useful for personalizing decision making.

	Clinical Decision
Decision 1	There is not sufficient evidence that the biomarker is useful for personalizing decision making
Decision 2	The biomarker may be useful for personalizing decision making, evidence suggests a treatment effect in subgroup 1
Decision 3	The biomarker may be useful for personalizing decision making, evidence suggests a treatment effect in subgroup 2
Decision 4	The biomarker may be useful for personalizing decision making, evidence suggests a qualitative interaction

Table 4.1: The decision space $\mathcal{D}$

4.2.3 Test Statistics

We utilize three test statistics to facilitate our two-stage procedure:

$$T_0 = \frac{\bar{X}_{1A} - \bar{X}_{1B} - \bar{X}_{2A} + \bar{X}_{2B}}{\hat{\sigma}\sqrt{\frac{4}{n_1} + \frac{4}{n_2}}} = \sqrt{p_2}T_1 - \sqrt{p_1}T_2$$

$$T_1 = \frac{\bar{X}_{1A} - \bar{X}_{1B}}{\hat{\sigma}\sqrt{\frac{4}{n_1}}}$$

$$T_2 = \frac{\bar{X}_{2A} - \bar{X}_{2B}}{\hat{\sigma}\sqrt{\frac{4}{n_2}}}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-4} \sum_{i\in\{1,2\},j\in\{A,B\},1\leq k\leq n_{ij}} (X_{ijk} - \bar{X}_{ij})^2$$

where $\bar{X}_{iT}$ is the sample average of subgroup i who are randomized to treatment T, $i \in \{1, 2\}$ and $T \in \{A, B\}$. These three are the standard test statistics used for testing the null hypothesis of no treatment effect in subgroup 1, the null hypothesis of no treatment effect in subgroup 2 and the null hypothesis of no treatment-subgroup interaction.

4.2.4 The Two-stage Decision Procedure

Our two-stage procedure is indexed by $M(c_0, c_1)$, where c_0 and c_1 are the critical values in stage I and stage II. The procedure is conducted as follows:

- In stage I, utilize test statistic T_0 and compare it with critical value $\pm c_0$. If $T_0 > c_0$ or $T_0 < -c_0$, proceed to stage II. Otherwise if $|T_0| \leq c_0$, stop the testing procedure and make clinical decision 1.
- In stage II, utilize test statistics T_1 and T_2 and compare them with critical value $\pm c_1$. Clinical decisions are made according to the decision rule specified in Table 2.

Stage I serves as gate keeper for the entire decision procedure since the existence of a quantitative interaction is the pre-requisite for a qualitative interaction. Ideas of stepwise gate keeping procedure have appeared in the literature, in particular with application to hypothesis testing in pharmaceutical science [21, 22]. Once the test statistics pass the gate keeper, the second stage serves to identify whether the "signs" of the treatment effects are consistent in the two groups. The signs of the treatment effects are consistent in the two subgroups if they are both significantly positive, significantly negative, or indistinguishable from 0. The decision is made according the consistency of the signs of subgroup treatment effects. In table 4.2 we partition the sample space into 10 parts and summarize the decision for each part. As examples, $|T_0| \leq c_0$ corresponds the part of the sample space with insufficient evidence for a quantitative interaction therefore we make decision 1. $T_0 < -c_0, T_1 < -c_1, |T_2| \leq c_1$ corresponds to the part of the sample space where patients in subgroup 1 benefits significantly more from treatment B while subgroup 2 patients show similar responses to both treatment. Here we make decision 2.

Decision Rule	Clinical Decision
$\lvert T_0\rvert \leq c_0$	Decision 1
$\lvert T_0\rvert > c_0, \lvert T_1\rvert \leq c_1, \lvert T_2\rvert \leq c_1$	Decision 1
$\lvert T_0\rvert > c_0, T_1 > c_1, T_2 > c_1$	Decision 1
$\lvert T_0\rvert > c_0, T_1 < -c_1, T_2 < -c_1$	Decision 1
$T_0 > c_0, T_1 > c_1, \lvert T_2\rvert \leq c_1$	Decision 2
$T_0 < -c_0, T_1 < -c_1, \lvert T_2\rvert \leq c_1$	Decision 2
$T_0 < -c_0, \lvert T_1\rvert \leq c_1, T_2 > c_1$	Decision 3
$T_0 > c_0, \lvert T_1\rvert \leq c_1, T_2 < -c_1$	Decision 3
$T_0 > c_0, T_1 > c_1, T_2 < -c_1$	Decision 4
$T_0 < -c_0, T_1 < -c_1, T_2 > c_1$	Decision 4

Table 4.2: The Decision Rule for the two-stage decision procedure for personalizing treatment

4.2.5 The Loss Function and Error probabilities

We specify a loss function $L(\theta, d)$ that is defined in Table 4.3. The rationale for choosing such loss function is the following. First, the loss function is 0 whenever decision 1 is reached. That is, we do not punish a false acceptance of the null hypothesis. Second, the loss function is 1 when any of decision 2, 3 or 4 is reached, if $\theta_1 = \theta_2$. That is, we punish a false rejection when there is no treatment-subgroup interaction. Third, in the case that θ belongs to the null space but a treatment-subgroup interaction exists, we punish the error of making decision 4, along with one of the decision 2 and 3. For example, suppose that the true parameter belongs to the region $\{\theta : \theta_1 > \theta_2 > 0\}$ and decision 2 is reached. Since $\theta_1 > \theta_2 > 0$, there is a positive treatment effect in both subgroups and subgroup 1 enjoys a larger treatment effect than subgroup 2. We argue that making decision 3 and 4 is a more severe error than making decision 2. The reason is that, the region $\{\theta : \theta_1 > \theta_2 > 0\}$ includes points $(\theta_1, \theta_2) = (K, \epsilon)$, where ϵ is infinitesimal small and K is large. For example, ϵ may be smaller than a small standardized effect size (0.2 in Cohen's benchmark), or any other clinical meaningful standardized effect size. Fourth, when θ belongs to the alternative space where there is no qualitative interaction, we punish the error of making decision 4 (restricted qualitative interaction), as well as the clinical decision that is incorrect in terms of the selecting the subgroup with a treatment effect. For example, the loss function is 1 for making decision 3 and 4 if the truth is $\theta_1 \neq 0, \theta_2 = 0$.

	Decision 1	Decision 2	Decision 3	Decision 4
$\theta_1 = \theta_2$	0	1	1	1
$\theta_1 > \theta_2 > 0$	0	0	1	1
$\theta_2 > \theta_1 > 0$	0	1	0	1
$\theta_1 < \theta_2 < 0$	0	0	1	1
$\theta_2 < \theta_1 < 0$	0	1	0	1
$\theta_1 \neq 0, \theta_2 = 0$	0	0	1	1
$\theta_1 = 0, \theta_2 \neq 0$	0	1	0	1
$\theta_1\theta_2 < 0$	0	0	0	0

Table 4.3: The loss function

4.3 Choosing the Critical Values c_0 and c_1

We follow the minimax paradigm and select c_0, c_1 to minimize the supreme of the risk function $R(\theta)$ over $\theta \in \mathcal{R}^2$. Simple calculation yields that, in order to control the supreme of the risk function, it is equivalent to control the following two expressions:

$$\sup_{\theta_1>\theta_2\geq 0} P_\theta(T_0 < -c_0, |T_1| \leq c_1, T_2 > c_1) + P_\theta(T_0 > c_0, |T_1| \leq c_1, T_2 < -c_1)$$
$$+P_\theta(T_0 > c_0, T_1 > c_1, T_2 < -c_1) + P_\theta(T_0 < -c_0, T_1 < -c_1, T_2 > c_1)$$

and

$$\sup_{\theta_1=\theta_2} P_\theta(T_0 > c_0, T_1 > c_1, |T_2| \leq c_1) + P_\theta(T_0 < -c_0, T_1 < -c_1, |T_2| \leq c_1)$$
$$+P_\theta(T_0 < -c_0, |T_1| \leq c_1, T_2 > c_1) + P_\theta(T_0 > c_0, |T_1| \leq c_1, T_2 < -c_1)$$
$$+P_\theta(T_0 > c_0, T_1 > c_1, T_2 < -c_1) + P_\theta(T_0 < -c_0, T_1 < -c_1, T_2 > c_1)$$

The next task is to search for the pair of (c_0, c_1) that controls the above error probability below level α. We use a sample of 5000 normally distributed random variables to approximate the error probabilities in the above two displays. We fix c_1 to be the critical value used in Gail and Simon's likelihood ratio test for qualitative interaction and perform a line search to find the smallest c_0 to control the total error rate below α. Table 4.4 summarizes the critical values at different subgroup percentages p_1, the proportion of subgroup 1 patients. The table shows that larger c_0 is required when the subgroup sample sizes are imbalanced.

p_1	c_0	c_1
0.10 or 0.90	2.06	1.64
0.20 or 0.80	2.05	1.64
0.30 or 0.70	1.96	1.64
0.40 or 0.60	1.96	1.64
0.50	1.96	1.64

Table 4.4: The critical values c_0 and c_1 at $\alpha = 0.05$

4.4 Comparing with Alternative Methods

Clinicians use a variety of subgroup analysis methods, in the current practice, to develop personalizing treatments. Here we briefly discuss three most commonly-encountered methods in the clinical trials literature and compare them with our proposed method.

Gail and Simon's likelihood ratio test of qualitative interaction [28]'s test of qualitative interaction can be used by clinicians who are interested in detecting qualitative interactions. Following Gail and Simon's paradigm, a biomarker is useful for personalizing decision making only if qualitative interaction exists. That is, there is sufficient evidence to demonstrate crossover treatment effects at different values of the biomarker. In contrast, according to the new definition, a biomarker is also useful for personalizing decision making when there is not sufficient evidence to recommend a particular treatment, at some value of the biomarker. The new definition matches the clinical practice by recognizing the increased patients' utility when they are given the freedom to choose a treatment based on their preferences. Our null hypothesis that the biomarker is not useful for personalizing decision making, is a proper subset of Gail and Simon's null hypothesis of no qualitative interaction. It follows that our alternative hypothesis properly contains Gail and Simon's alternative hypothesis of qualitative interaction by including $\{\theta : \theta_1 = 0, \theta_2 \neq 0\}$ and $\{\theta : \theta_1 \neq 0, \theta_2 = 0\}$. The proposed procedure, not only is capable to detect the restricted qualitative interactions, but is also capable of detecting the generalized qualitative interactions which are also informative for personalizing decision making (clinical decision 2 and 3).

If one puts aside the differences of the underlying hypotheses and focus only on the intersection region of the two alternative hypotheses ($\theta_1\theta_2 < 0$), it is desirable that a proposed procedure has at least as much the power to detect qualitative interactions as Gail and Simon's test. In other words, the procedure should not sacrifice the power of making

clinical decision 4 in the presence of clinical decision 2 and 3. It is straight forward to verify that, asour procedure has the same power of detecting qualitative interactions as Gail and Simon's test. Recall that the second stage critical value in our decision procedure is the critical value used by Gail and Simon. Our proposed procedure has the same power in detecting a restricted qualitative interaction if

$$\begin{aligned}\{(T_1,T_2):\sqrt{p_2}T_1-\sqrt{p_1}T_2>c_0,T_1>c_1,T_2<-c_1\}\\ =\{(T_1,T_2):T_1>c_1,T_2<-c_1\}\\ \{(T_1,T_2):\sqrt{p_2}T_1-\sqrt{p_1}T_2<-c_0,T_1<-c_1,T_2>c_1\}\\ =\{(T_1,T_2):T_1<-c_1,T_2>c_1\}\end{aligned}$$

One can easily verify, based on table 4.4 that the above two equality holds if $0.1 \leq p_1 \leq 0.9$. In other words, our procedure has the same power in detecting a restricted qualitative interaction when the subgroup sample size imbalance is not extreme; otherwise the proposed procedure has inferior power to detect a qualitative interaction.

Subgroup analysis which tests subgroup hypotheses

[64] summarized some of the current practices in subgroup analysis. The summary pointed out that a lot of subgroup analysis (37% of the reports in their survey) has been conducted by simply testing the subgroup treatment effects in each subgroup. In the context of two subgroups, the two subgroup hypotheses are $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$. Decisions are thus made based on the p-values as well as the signs of the test statistics. For example, if the one-sided p-value associated with T_1 is less than 0.025 and $T_1 > 0$ while the p-value associated with T_2 is greater than 0.025, clinical decision 2 may be reached. This procedure, however, cannot proper control the errors in table 4.3. In fact, a simple simulation shows that, the error probability of making clinical decision 2 or 3 is approximately 0.25 when $\theta_1 = \theta_2 = 2$, if both H_1 and H_2 are tested at level 0.05. The reason is that this procedure analyze treatment effects in each subgroup separately while no effort has been taken in analyzing the treatment-subgroup interaction. The inflated error probability at $\theta_1 = \theta_2 = 2$ is simply due to the sum of the Type II errors.

Subgroup analysis which tests treatment-subgroup interaction, as well as subgroup hypotheses

A more principled way of conducting subgroup analysis (for example, [17]) is to jointly test the hypothesis of treatment-subgroup interaction, as well as the hypotheses concerning treatment effects in each subgroup. In the context of two subgroups, the three hypotheses are $H_0 : \theta_1 - \theta_2 = 0$, $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$. One may control the familywise error rate using Bonferroni adjustment or any other multiple testing procedure (for example,

Holm 1979). When Bonferroni adjustment is used, the level of each individual hypothesis is $\alpha/3$. This procedure properly controls the error probabilities we proposed. The Bonferroni adjustment itself, however, may result in using conservative critical values in our specific problem. Our proposed procedure suggests using critical values $c_0 = \sqrt{2}z_{\alpha/2}$ and $c_1 = z_\alpha$. Therefore, the stage I of the proposed procedure is equivalent to testing H_0 at level α, while stage II is equivalent to testing H_1 and H_2 at level 2α. Using unnecessarily large critical values leads to undermined power in rejecting the null hypothesis H and making clinical conclusion 2, 3 and 4.

The future work of this project includes

1. Generalize the framework and decision procedure to the setting where the biomarker takes more than two values. This problem is more challenging than the binary biomarker problem we consider due to the increased complexity of the decision space and different types of errors.

2. Non-normal data distributions, unknown variances, multiple regression. The normality assumption we made in deriving the procedure may be an oversimplification of the reality. It is desirable to extend the current procedure to more general data distributions, possible with unknown variances. To obtain more precise estimates of θ_1 and θ_2, a commonly-adopted method is to use regression model to adjust for the heterogeneity in other baseline variables. There, the estimators of θ_1 and θ_2 and thus the test statistics T_1 and T_2 will be correlated. Deriving critical values in these more general setting may require bootstrapping.

3. Multiple decision point. In treating chronic diseases, treatment assignments usually need to be adjusted over time according to the changing needs and performances of the patients. It is desirable, as a consequence, to personalize decision making over the entire course of the treatment. Our ultimate goal is to construct hypothesis testing procedure for personalizing treatment at multiple decision points, utilizing data from the sequential multiple assignment randomized trials ([56]).

BIBLIOGRAPHY

[1] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012.

[2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[4] Stephanie Bauer, Judith de Niet, Reinier Timman, and Hans Kordy. Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight. *Patient education and counseling*, 79(3):315–319, 2010.

[5] R.L. Berger. Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association*, 84(405):192–199, 1989.

[6] R.L. Berger and D.D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

[7] Dimitri P Bertsekas. Nonlinear programming. 1999.

[8] Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[9] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[10] Karen L Bierman, Robert L Nix, Jerry J Maples, and Susan A Murphy. Examining clinical judgment in an adaptive intervention design: The fast track program. *Journal of Consulting and Clinical Psychology*, 74(3):468, 2006.

[11] Patrick Billingsley. The lindeberg-levy theorem for martingales. *Proceedings of the American Mathematical Society*, 12(5):788–792, 1961.

[12] Vivek S Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

[13] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[14] T. Cai, L. Tian, P.H. Wong, and LJ Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011.

[15] Marco C Campi and Simone Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011.

[16] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[17] J. Cohen and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum, 1975.

[18] Linda M Collins, Susan A Murphy, and Karen L Bierman. A conceptual framework for adaptive preventive interventions. *Prevention science*, 5(3):185–196, 2004.

[19] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1797–1806. ACM, 2008.

[20] Walter Dempsey, Peng Liao, Pedja Klasnja, Inbal Nahum-Shani, and Susan A Murphy. Randomised trials for the fitbit generation. *Significance*, 12(6):20–23, 2015.

[21] Alex Dmitrienko, Ajit C Tamhane, Xin Wang, and Xun Chen. Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal*, 48(6):984–991, 2006.

[22] Alex Dmitrienko, Ajit C Tamhane, and Brian L Wiens. General multistage gatekeeping procedures. *Biometrical Journal*, 50(5):667–677, 2008.

[23] Leonard H Epstein, Katelyn A Carr, Meghan D Cavanaugh, Rocco A Paluch, and Mark E Bouton. Long-term habituation to food in obese and nonobese women. *The American journal of clinical nutrition*, 94(2):371–376, 2011.

[24] Leonard H Epstein, Jennifer L Temple, James N Roemmich, and Mark E Bouton. Habituation as a determinant of human food intake. *Psychological review*, 116(2):384, 2009.

[25] Anthony V Fiacco and Yo Ishizuka. Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, 27(1):215–235, 1990.

[26] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[27] H. Finner and K. Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Annals of statistics*, pages 1194–1213, 2002.

[28] M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372, 1985.

[29] J.J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6):3782–3810, 2010.

[30] L. Gunter, J. Zhu, and SA Murphy. Variable selection for qualitative interactions. *Statistical methodology*, 8(1):42–55, 2011.

[31] David H Gustafson, Bret R Shaw, Andrew Isham, Timothy Baker, Michael G Boyle, and Michael Levy. Explicating an evidence-based, theoretically informed, mobile technology-based system to improve outcomes for people in recovery for alcohol dependence. *Substance use & misuse*, 46(1):96–111, 2011.

[32] P. Hansen. Asymptotic tests of composite hypotheses. *Brown University Economics Working Paper*, 2003.

[33] Thomas P Hayes. A large-deviation inequality for vector-valued martingales.

[34] Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The Dependence of Effective Planning Horizon on Model Accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 1181–1189, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems.

[35] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2-3):193–208, 2002.

[36] R.W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010.

[37] Abby C King, Eric B Hekler, Lauren A Grieco, Sandra J Winter, Jylana L Sheats, Matthew P Buman, Banny Banerjee, Thomas N Robinson, and Jesse Cirimele. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PloS one*, 8(4):e62613, 2013.

[38] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

[39] Predrag Klasnja, Eric B. Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A. Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl):1220–1228, 2015.

[40] Levente Kocsis and Csaba Szepesvri. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293. Springer, 2006.

[41] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *NIPS*, volume 13, pages 1008–1014, 1999.

[42] Ricardo Lage, Ludovic Denoyer, Patrick Gallinari, and Peter Dolog. Choosing which message to publish on social networks: a contextual bandit approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 620–627. IEEE, 2013.

[43] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[44] Philip W Lavori and Ree Dawson. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

[45] Philip W Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20, 2004.

[46] Philip W Lavori, Ree Dawson, and A John Rush. Flexible treatment strategies in chronic disease: clinical and research implications. *Biological Psychiatry*, 48(6):605–614, 2000.

[47] E.L. Lehmann. A theory of some multiple decision problems, i. *The Annals of Mathematical Statistics*, pages 1–25, 1957.

[48] E.L. Lehmann. A theory of some multiple decision problems. ii. *The Annals of Mathematical Statistics*, 28(3):547–572, 1957.

[49] E.L. Lehmann and J.P. Romano. *Testing statistical hypotheses*. Springer, 2005.

[50] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[51] O. Linton, K. Song, and Y.J. Whang. An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, 154(2):186–202, 2010.

[52] Jared K Lunceford, Marie Davidian, and Anastasios A Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

[53] R. Marcus, P. Eric, and K.R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[54] Douglas B Marlowe, David S Festinger, Patricia L Arabia, Karen L Dugosh, Kathleen M Benasutti, Jason R Croft, and James R McKay. Adaptive interventions in drug court a pilot experiment. *Criminal Justice Review*, 33(3):343–360, 2008.

[55] James R McKay. Continuing care research: What we have learned and where we are going. *Journal of substance abuse treatment*, 36(2):131–145, 2009.

[56] Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005.

[57] Susan A Murphy, MJ Van Der Laan, and James M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[58] Inbal Nahum-Shani, Shawna N Smith, Ambuj Tewari, Katie Witkiewitz, Linda M Collins, Bonnie Spring, and S Murphy. Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. *Methodology Center technical report*, (14-126), 2014.

[59] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

[60] Kevin Patrick, Fred Raab, Marc Adams, Lindsay Dillon, Marion Zabinski, Cheryl Rock, William Griswold, and Gregory Norman. A text message-based intervention for weight loss: randomized controlled trial. *Journal of medical Internet research*, 11(1):e1, 2009.

[61] Vianney Perchet, Philippe Rigollet, et al. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

[62] M.D. Perlman and L. Wu. The emperors new tests. *Statistical Science*, 14(4):355–369, 1999.

[63] S. Piantadosi and MH Gail. A comparison of the power of two tests for qualitative interactions. *Statistics in medicine*, 12(13):1239–1248, 1993.

[64] S.J. Pocock, S.E. Assmann, L.E. Enos, and L.E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

[65] Mashfiqui Rabbi, Angela Pfammatter, Mi Zhang, Bonnie Spring, and Tanzeem Choudhury. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth and uHealth*, 3(2):e42, 2015.

[66] Hollie A Raynor and Leonard H Epstein. Dietary variety, energy regulation, and obesity. *Psychological bulletin*, 127(3):325, 2001.

[67] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

[68] William T Riley, Daniel E Rivera, Audie A Atienza, Wendy Nilsen, Susannah M Allison, and Robin Mermelstein. Health behavior models in the age of mobile interventions: are our theories up to the task? *Translational behavioral medicine*, 1(1):53–71, 2011.

[69] Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[70] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period?application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[71] E. Russek-Cohen and R.M. Simon. Evaluating treatments when a gender by treatment interaction may exist. *Statistics in medicine*, 16(4):455–464, 1997.

[72] Christy K Scott and Michael L Dennis. Results from two randomized clinical trials evaluating the impact of quarterly recovery management checkups with adult chronic substance users. *Addiction*, 104(6):959–971, 2009.

[73] R. Simon. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in medicine*, 21(19):2909–2916, 2002.

[74] E. Sonnemann and H. Finner. Vollsta?ndigkeitssa?tze fu?r multiple testprobleme. *Multiple Hypothesenpru?fung*, pages 121–135, 1988.

[75] Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

[76] Brian Suffoletto, Clifton Callaway, Jeff Kristan, Kevin Kraemer, and Duncan B Clark. Text-message-based drinking assessments and brief interventions for young adults discharged from the emergency department. *Alcoholism: Clinical and Experimental Research*, 36(3):552–560, 2012.

[77] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063, 1999.

[78] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

[79] S Armagan Tarim, Suresh Manandhar, and Toby Walsh. Stochastic constraint programming: A scenario-based approach. *Constraints*, 11(1):53–80, 2006.

[80] Peter F Thall, Hsi-Guang Sung, and Elihu H Estey. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 2011.

[81] Peter F Thall and J Kyle Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in medicine*, 24(13):1947–1964, 2005.

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

[83] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[84] Kyriakos G Vamvoudakis and Frank L Lewis. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5):878–888, 2010.

[85] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[86] Abdus S Wahed and Anastasios A Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

[87] Abdus S Wahed and Anastasios A Tsiatis. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93(1):163–177, 2006.

[88] Toby Walsh. Stochastic constraint programming. In *ECAI*, volume 2, pages 111–115, 2002.

[89] Katie Witkiewitz, Sruti A Desai, Sarah Bowen, Barbara C Leigh, Megan Kirouac, and Mary E Larimer. Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. *Psychology of Addictive Behaviors*, 28(3):639, 2014.

[90] Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

[91] Jeremy Wyatt. Exploration and inference in learning from reinforcement. 1998.

[92] Yuhong Yang, Dan Zhu, et al. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

[93] D. Zelterman. On tests for qualitative interactions. *Statistics & probability letters*, 10(1):59–63, 1990.