Just-In-Time Adaptive Interventions: Experiment, Inference and Online Learning

by

Peng Liao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2019

Doctoral Committee:

Emeritus Professor Susan A. Murphy, Co-Chair
Associate Professor Ambuj Tewari, Co-Chair
Research Associate Professor Daniel Almirall
Assistant Professor Predrag Klasnja

Peng Liao

pengliao@umich.edu

ORCID iD: 0000-0002-5854-0515

© Peng Liao 2019

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

The use and development of mobile interventions are experiencing rapid growth. Ideally, mobile devices can be used to provide treatment/support whenever needed and to adapt treatment to the context of the user. Just-In-time Adaptive Interventions (JITAIs) are composed of decision rules that map a users context (e.g., user's behaviors, location, current time, social activity, stress and urges to smoke) to a treatment that is delivered to the user via the mobile device in near real-time. Advancements in mobile health engineering and technology (e.g., passive stress sensing) continue to bring us closer to being able to provide interventions in this way. However, a number of important gaps in data science must be addressed before mobile devices can be used to deliver on the promise of JITAIs. First, there is a need for experimental designs to collect data that can be used to assess the effectiveness of the sequence of treatments delivered by a mobile device on health outcomes in order to support the development of JITAIs. Second, there is a need for data-driven methods to inform the construction of efficacious JITAIs. In the vast majority of currently deployed JITAIs, the decision rules underpinning JITAIs are formulated using domain expertise and clinical experience, with very limited use of data evidence.

In this dissertation, we make several contributions by tackling the above- mentioned data science barriers to effective JITAI development in mobile health. First, we propose a micro-randomized trial (MRT) design and develop the primary analysis for assessing the proximal causal effect of treatments. In addition, we develop stratified micro-randomized trials for the setting where there is a time-varying, discrete variable, and the primary analysis focuses on how the effectiveness of interventions changes with this variable. We also develop a novel algorithm to design the

randomization scheme for this setting when there is an average constraint on the number of times interventions that should be sent in a certain time interval. Second, we develop a semi-parametric model to estimate the long-term average of health outcomes that would accrue should a given JITAI be followed. We derive the rate of convergence and the asymptotic normality of the proposed estimator. Third, we develop an online learning algorithm that continuously learns and improves the JITAI as the data is collected from the user. The proposed algorithm introduces a proxy of future outcomes based on a dosage variable to capture the delayed effect of sending the interventions due to the treatment burden.

CHAPTER 1

Introduction

Due to recent advances in mobile technologies such as smartphone along with sophisticated wearable sensors, mobile health (mHealth) technologies are drawing much attention in the behavioral health communities. For example, wearable sensors for detecting physical activity and physiological states are now widely available. Modern smartphones could sense their users environment in real time and mine data from their calendars, email, and other applications. This allows scientists for the first time to not only unobtrusively collect real-time data, but also deliver interventions at moments when they can most readily influence a persons behavior. The potential of mHealth interventions may be best realized when they could adapt to individuals response and context, and deliver effective intervention options at the right time and location, which gives rise to the concept of Just-in-time Adaptive Intervention.

Just-In-Time Adaptive Interventions (JITAIs) aim to use real-time, either passively or actively collected, information on the user to deliver the right intervention component at the right time and location to optimally prevent negative health outcomes and promote the adoption and maintenance of healthy behaviors. These could be operationalized via decision rules (or treatment policies) which map the users context (weather, location, current time, social activity, stress, urges to smoke, etc.) to intervention options delivered via a mobile device, specifying whether, when, and what type of intervention should be provided. These treatment policies are being used to intervene in physical activity, eating disorders, alcohol use, mental illness, obesity/weight management, and other chronic disorders.

Despite the technological advances to realize JITAIs, there are barriers to the development of effective JITAIs. *First, researchers currently do not have the appropriate experimental design to gather data/evidence to decide whether or not those mobile interventions have impact on users healthy behaviors and to support the construction of JITAIs.* Commonly used experimental designs are not sufficient to support development of just-in-time interventions because they do not enable researchers to determine empirically when a particular intervention component should be delivered

and whether a just-in-time intervention that was delivered had the intended effect. *Second, theres a lack of data-based methods to inform the construction of efficacious evidence-based JITAIs.* In the applications of JITAIs mentioned earlier and throughout much of mobile health, the treatment policies are formulated using domain expertise and clinical experience. Researchers have recently argued that the extent to which our behavioral theories can guide the development of just-in-time interventions is limited. In particular, although intervention components included in a JITAI are often based on behavioral theories, with rare exceptions these theories are mature enough to specify dynamics of human behavior to guide the design of the decision rules that precisely specify when particular intervention components should be delivered in order to ensure the interventions have the intended effects and optimize the long-term efficacy of the interventions.

In this dissertation, we make several contributions by tackling the above- mentioned data science barriers to effective JITAI development in mobile health. *The first contribution is on the development of new experimental designs for testing the effectiveness of the mobile interventions.* In Chapter 2, we propose a micro-randomized trial design, where treatments are sequentially randomized throughout the conduct of the study. We use the potential outcome framework to define the causal treatment effect in this setting and propose a test statistics as well as a sample size calculator to help scientists designing a micro-randomized trial. This work has been published in *Statistics in Medicine*. [1]. In Chapter 3, we develop stratified micro-randomized Trial, a generalization of micro-randomized trial for the setting where there is a time-varying, discrete variable, and the scientific interest is to understand how the effectiveness of interventions changes with this variable. This is joint work with Walter Dempsey and others and is currently under review for *Annals of Applied Statistics*. The chapter 3 is adapted from the preprint version [2]. To design stratified micro-randomized trials, it is crucial to ensure randomization occurs sufficiently enough at each level of the time-varying variable. However, treatment burden consideration often imposes a constraint on the number of times the intervention should be delivered. In Chapter 4, we develop a novel algorithm that determines the randomization probability to satisfy the average constraint and uniformly spread across time. This work has been published in Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) [3].

The second contribution is that we develop a data analysis method to assess the quality of JITAIs of interest, using data collected from the above-mentioned trial design. In Chapter 5, we model the decision-making problem using Markov Decision Process framework and develop a semi-parametric model to estimate the long-term average of the health outcomes that would accrue should a given JITAI be followed. We use a flexible function class to model the relative value function and derive asymptotic theory on the consistency and asymptotic normality of the

estimated average reward.

The third contribution is on developing an online Reinforcement Learning (RL) algorithm for mobile health to continuously learn and improve the treatment policy embedded in the JITAI as the data is collected from the person. Reinforcement Learning (RL) is an area of Machine Learning in which an algorithm learns how to act optimally by continuously interacting with the unknown environment. There are many existing online RL algorithms for automatic optimization of action sequences in RL literature. However, many challenges remain that need to be carefully addressed before RL can be usefully deployed to adapt and optimize mobile health interventions. One key challenge arising in mobile health is that the RL algorithm should learn quickly. Most online RL algorithms require the agent to interact many times with the environment prior to performing well. This is impractical in mobile health applications as users can lose interest and disengage quickly. On the other hand, the RL algorithm must adjust for long-term effects of current actions. In mobile health, interventions often tend to have a positive effect on the immediate reward, but likely produce a negative impact on future rewards due to user habituation and/or burden. To address these two challenges, in Chapter 6, we develop an algorithm that mixes between Thompson Sampling Bandit and full RL algorithm.

CHAPTER 2

Micro-Randomized Trails

The use and development of mobile interventions are experiencing rapid growth. Mobile interventions are used across the health fields and include treatments to improve HIV medication adherence [4, 5], to increase activity [6], supplement counseling/pharmacotherapy in treatment for substance use [7, 8], reinforce abstinence in addictions [9, 10] and to support recovery from alcohol dependence [11, 12]. Mobile interventions for adherence to anti-retroviral therapy and smoking cessation have shown sufficient effectiveness and replicability in trials and have been recommended for inclusion in health services [13].

However, as Nilsen *et al.* [14] state, "In fact, the development of mHealth technologies is currently progressing at a much faster pace than the science to evaluate their validity and efficacy, introducing the risk that ineffective or even potentially harmful or iatrogenic applications will be implemented."Indeed reviews, while reporting preliminary evidence of effectiveness, call for more programmatic, data-based approaches to constructing mobile interventions [13, 15]. In particular, these reviews call for research that focuses on data-informed development of these complex multicomponent interventions prior to their evaluation in standard randomized controlled trials. But methods for using data to inform the design and evaluation of adaptive mobile interventions have lagged behind the use and deployment of these interventions [14, 16, 17].

Many mobile interventions are designed to be "just-in-time" interventions, meaning that they intend to provide treatments that help an individual make healthy decisions in the moment, such as engaging in a desirable behavior (e.g., taking a medication on time) or effectively coping with a stressful situation. As such, mobile interventions are often intended to have proximal, near-term effects. A first approach toward developing data-based methods for evaluation of mobile health interventions is to provide an experimental design for testing the proximal effects of the treatments.

In this chapter, we introduce micro-randomized trial design for this purpose. In a micro-randomized trial, treatments are sequentially randomized throughout the conduct of the study, with the result that each participant may be randomized at the hundreds or thousands of occasions at

which a treatment might be provided. This repeated randomization of treatments under investigation enables causal modeling of each treatment's time-varying proximal effect as well as modeling of time-varying effect moderation. Thus, the micro-randomized trial can be seen as a first experimental step in the development of effective mobile interventions that are composed of sequences of treatments. We propose to size the trial to detect the proximal main effect of the treatments. This is akin to the use of factorial designs for use in constructing multi-component interventions. In these factorial designs [18, 19], a first analysis often involves testing if the main effect of each treatment is equal to 0. This work is motivated by our collaboration on the HeartSteps mobile application for increasing physical activity, which we will use to illustrate our discussion. In the following section, we briefly introduce HeartSteps. In section 2.2 and 2.3, we introduce the micro-randomized trials design and precisely define the proximal main effect of a treatment, using the language of potential outcomes. We develop the test statistic for assessing the proximal effect of a treatment as well as an associated sample size calculator in section 2.4 and 2.5. Next we provide simulation evaluation of the sample size calculator. We end, in Section 2.7, with a discussion.

2.1 HeartSteps: Physical Activity Study

HeartSteps is a mobile health study focused on promoting physical activity among sedentary individuals. One of the intervention components in HeartSteps is suggestions for physical activity which are tailored to the person's current context. HeartSteps can deliver these suggestions at any of the five time intervals during the day, which correspond roughly to morning commute, mid-day, mid-afternoon, evening commute, and post-dinner times. When a suggestion is delivered, the user's phone plays a notification sound, vibrates and lights up, and the suggestion is displayed on the lock screen of the phone. These suggestions encourage activity in the current context and are intended to have an effect in the near future , e.g., getting a person to walk.

2.2 Micro-Randomized Trial

In general an individual's longitudinal data, recorded via mobile devices that sense and provide treatments, can be written as

$$\{O_0, O_1, A_1, O_2, A_2, \ldots, O_t, A_t, \ldots, O_T, A_T, O_{T+1}\}$$

where, t indexes decision times, O_0 is a vector of baseline information (gender, ethnicity, etc.) and $O_t(t \geq 1)$ is information collected between time $t-1$ and t (e.g., summary measures of recent activity levels, engagement, and burden; day of week; weather; busyness indicated by smartphone calendar, etc.). The treatment at time t is denoted by A_t; throughout this chapter we consider binary options for the treatments (e.g., the treatment is on or off). The proximal response, denoted by Y_{t+1}, is a known function of $\{O_t, A_t, O_{t+1}\}$. Here we assume that the longitudinal data are independent and identically distributed across N individuals. Note that this assumption would be violated, if for example, some of the treatments are used to enhance social support between individuals in the study.

In HeartSteps, data (O_t) is collected both passively via sensors and via participant self-report. Each participant is provided a "Jawbone" band, worn at the wrist, which collects daily step count and the amount of sleep the user had the previous night. Furthermore sensors on the phone are used to collect a variety of information at each of the 5 time points during the day, including the time-stamp, location, busyness of planned activities on the phone calendar and other activity on the phone. Each evening, self-report data is collected including utility and burden ratings. The proximal response, Y_{t+1}, for activity suggestions is the step count collected in the next 30 minutes following time t.

A decision time is a point in time at which—based on participant's current state, past behavior, or current context—treatment may need to be delivered. Decision times vary by the nature of the intervention component. In HeartSteps, the decision times for activity suggestions are 5 times per day over the 42 day study duration. For an alcohol-recovery application that provides an intervention when an individual goes within 10 feet of a high risk location (e.g., a liquor store), decision points might be every 1 minute, the frequency at which the application would get the person's current location and assess whether she is close to a high-risk location. In a long-term study of an intervention for multiple health behaviors, the decision points might be weekly or monthly at which times, decisions are made regarding whether to change the focus from one behavior (e.g., physical activity) to another (e.g., diet). Finally, in many studies there is an option for an individual to press a "panic"button, indicating the need for help; for such interventions, decision times correspond to times at which the panic button is pressed.

A micro-randomized trial is a trial in which at each decision time t, participants are randomized to a treatment option, denoted by A_t. Treatment options may correspond to whether or not a treatment is provided at a decision time; for example in HeartSteps, whether or not the individual is provided a lock-screen activity suggestion. Or treatment options may be alternative types of treatment that can be provided at the same decision time; for example, a daily step goal treatment might

have two options, a fixed 10,000-steps-a-day goal or an adaptive goal based on the user's activity level on the previous day. Considerations of treatment burden often imply that the randomization will not be uniform. For example in HeartSteps, the randomization probability is 0.4, so that, if an individual is always available, on average 2 lock-screen activity messages are delivered per day.

In designing, that is, determining the sample size for, a micro-randomized trial we focus on the reduced longitudinal data

$$\{I_1, A_1, Y_2, I_2, A_2, Y_3, \ldots, I_t, A_t, Y_{t+1}, \ldots, I_T, A_T, Y_{T+1}\}.$$

The variable, I_t is an "availability"indicator. The availability indicator is coded as $I_t = 1$ if the individual is available for treatment and $I_t = 0$ otherwise. At some decision times feasibility, ethics or burden considerations mean that the individual is unavailable for treatment and thus A_t should not be delivered. Consider again HeartSteps: if sensors indicate that the individual is likely driving a car or the individual is currently walking, then the lock-screen activity message should not be sent. Other examples of when individuals are unavailable for treatment include: in the alcohol recovery setting, an "warning"treatment would only be potentially provided when sensors indicate that the individual is within 10 feet of a high risk location or a treatment might only be provided if the individual reports a high level of craving. If the application has a panic button, then only in an x second interval in which the panic button is pressed is it appropriate to provide "panic button"treatments. Individuals may be unavailable for treatment by choice. For example, the HeartSteps application permits the individual to turn off the lock-screen activity messages; this option is considered critical to maintaining participant buy-in and engagement with HeartSteps. After viewing the lock-screen activity message, the individual has the option of turning off the lock-screen messages for 4 , 8 or 12 hours. After the specified time interval, the delivery of lock-screen messages automatically turns on again. To summarize, the availability indicator at time t is the indicator for the subpopulation at time t among which we are interested in assessing the proximal main effect of the treatment; *we are uninterested in assessing the proximal main effect of a treatment among individuals for whom it is unethical to provide treatment or for whom it makes no scientific sense to provide treatment or among those who refuse to be provided a treatment.*

2.3 Proximal Main Effect of Treatment

As discussed above, treatments in mobile health interventions are often designed so as to have a proximal effect (e.g., increase activity in near future, help an individual manage current cravings

for drugs or food, take medications on schedule, etc.). As a result, a first question in developing a mobile health intervention is whether the treatments have a proximal effect. Here we develop sample size formulae that guarantee a stated power to detect the proximal effect of a treatment. In particular we aim to test if the proximal main effect is zero.

To define the proximal main effect of a treatment, we use potential outcomes [20, 21, 22]. Our use of potential outcome notation is slightly more complicated than usual because treatment can only be provided when an individual is available. As a result, we index the potential outcomes by decision rules that incorporate availability. In particular define $d(a,i)$ for $a \in \{0,1\},\ i \in \{0,1\}$ by $d(a,0) =$"unavailable-do nothing"and $d(a,1) = a$. Then for each $a_1 \in \mathcal{A}_1 = \{0,1\}$, define $D_1(a_1) = d(a_1, I_1)$. Then we denote the potential proximal responses following decision time 1 by $\{Y_2^{D_1(1)},\ Y_2^{D_1(0)}\}$ and denote the potential availability indicators at decision time 2 by $\{I_2^{D_1(1)},\ I_2^{D_1(0)}\}$. Next for each $\bar{a}_2 = (a_1, a_2)$ with $a_1, a_2 \in \{0,1\}$, define $D_2(\bar{a}_2) = d(a_2, I_2^{D_1(a_1)})$. Define $\bar{D}_2(\bar{a}_2) = (D_1(a_1), D_2(\bar{a}_2))$. A potential proximal response following decision time 2 and corresponding to $\bar{a}_2$ is $Y_3^{\bar{D}_2(\bar{a}_2)}$ and a potential availability indicator at decision time 3 is $I_3^{\bar{D}_2(\bar{a}_2)}$. Similarly, for each $\bar{a}_t = (a_1, \ldots, a_t) \in \mathcal{A}_t = \{(a_1, \ldots, a_t) \big| a_i \in \{0,1\}, i = 1, \ldots, t\}$, define $D_t(\bar{a}_t) = d(a_t, I_t^{\bar{D}_{t-1}(\bar{a}_{t-1})})$ and $\bar{D}_t(\bar{a}_t) = (D_1(a_1), \ldots, D_t(\bar{a}_t))$. For each $\bar{a}_t = (a_1, \ldots, a_t) \in \mathcal{A}_t$, the potential proximal response is $Y_t^{\bar{D}_{t-1}(\bar{a}_{t-1})}$ (following decision time $t-1$) and potential availability indicator is $I_t^{\bar{D}_{t-1}(\bar{a}_{t-1})}$ at decision time t.

We define the proximal main effect of a treatment at time t among available individuals by:

$$\beta(t) = \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},1)} - Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},0)} \mid I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1\big]$$

where the expectation is taken with respect to the distribution of the potential outcomes and randomization in $\bar{A}_{t-1}$. This proximal effect is conditional in that the effect of treatment at time t is defined for only individuals available for treatment at time t, that is, $I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1$. This proximal effect is a main effect in that the effect is marginal over any effects of $\bar{A}_{t-1}$. The former conditional aspect of the definition is related to the concept of viable or feasible dynamic treatment regimes [23, 24] in which one assesses only the causal effect of treatments that can actually be provided.

Consider the proximal main effect, $\beta(t)$, as t varies across time. $\beta(t)$ may vary across time for a variety of reasons. To see this consider the case of HeartSteps. Here $\beta(t)$ might initially increase with increasing t as participants learn and practice the activities suggested on the lock-screen. For larger t one might expect to see decreasing or flat $\beta(t)$ due to habituation (participants begin to, at least partially, ignore the messages). This time variation in $\beta(t)$ can be attributed to both the immediate effect of a lock-screen activity message as well as interactions between the past lock-screen activity messages and the present activity message; the time variation occurs at

least partially due to the marginal character of $\beta(t)$. Alternately the conditional definition of $\beta(t)$ means that the effect is only defined among the population of individuals who are available at decision time t. Changes in this population may cause changes in $\beta(t)$ across time. Again consider HeartSteps. At earlier time points, participants may be highly engaged, yet have not developed habits that in various ways increase their activity, thus most participants will be available. However as time progresses, some participants may develop sufficiently positive activity habits or anticipate activity suggestions, thus at later decision times these participants may be already active and thus unavailable to receive a suggestion. Other participants may become increasing disengaged and repeatedly turn off the lock-screen activity messages; these participants are also unavailable. Thus as time progresses, $\beta(t)$ may vary due to the subpopulation of participants among whom it is appropriate to assess the effect of the lock-screen activity messages.

Our main objective in determining the sample size will be to assure sufficient power to detect alternatives to the null hypothesis of no proximal main effect, $\mathrm{H}_0 : \beta(t) = 0,\ t = 1, \ldots T$ for a trial with T decision points (if $\beta(t)$ is nonzero then for the population available at decision time t, there is a proximal effect). The proposed test will be focused on detecting smooth, i.e., continuous in t, alternatives to this null hypothesis.

To express $\beta(t)$ in terms of the observed data distribution, we assume consistency [21, 22]. This assumption is that for each t, the observed Y_t and observed I_t equal the corresponding potential outcomes, $Y_t^{\bar{D}_{t-1}(\bar{a}_{t-1})}$, $I_t^{\bar{D}_{t-1}(\bar{a}_{t-1})}$ whenever $\bar{A}_{t-1} = \bar{a}_{t-1}$. This assumption may be violated if some of the treatments promote social linkages between participants, for example, to enhance social/emotional support or to compete in mobile games. In these cases it would be more appropriate to additionally index each individual's potential outcomes by other participants' treatments. The micro-randomization plus the consistency assumption implies that the proximal main effect of treatment at time t among available individuals, $\beta(t)$ can be written as,

$$
\begin{aligned}
\beta(t) &= \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},1)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1\big] - \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},0)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1\big] \\
&= \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},1)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, A_t = 1\big] - \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_{t-1},0)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, A_t = 0\big] \\
&= \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_t)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, A_t = 1\big] - \mathbb{E}\big[Y_{t+1}^{\bar{D}_t(\bar{A}_t)} \big| I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, A_t = 0\big] \\
&= \mathbb{E}[Y_{t+1} | I_t = 1, A_t = 1] - \mathbb{E}[Y_{t+1} | I_t = 1, A_t = 0]
\end{aligned}
$$

where the second equality follows from the randomization of the A_t's and the last equality follows from the consistency assumption.

2.4 Test Statistic

Our sample size formula is based on a test statistic for use in testing $\mathbf{H}_0 : \beta(t) = 0,\ t = 1, \ldots T$ against a scientifically plausible alternative. This alternative should be formed based on conversations with domain experts. Here we construct a test statistic to detect alternatives that are, at least approximately, linear in a vector parameter, β, that is, alternatives of the form $Z_t^\top \beta$, where the $p \times 1$ vector, Z_t, is a function of t and covariates that are unaffected by treatment such as time of day or day of week. In the case of HeartSteps, a plausible alternative is quadratic:

$$Z_t^\top \beta = \left(1, \left\lfloor \frac{t-1}{5} \right\rfloor, (\left\lfloor \frac{t-1}{5} \right\rfloor)^2\right)\beta \tag{2.1}$$

where $\beta = (\beta_1, \beta_2, \beta_3)^\top$ ($p = 3$). Recall that in HeartSteps there are 5 decision times per day; $\lfloor \frac{t-1}{5} \rfloor$ translates decision times t to days. This rather simplistic parametrization marginalizes across the day and treats the weekends and weekdays similarly.

We propose to use the alternate, $\mathbf{H}_1 : \beta(t) = Z_t^\top \beta$, $t = 1, \ldots, T$ to construct the test statistic. We base the test statistic on the estimator of β in a least squares fit of a working model. A simple working model based on the alternative is:

$$\mathbb{E}[Y_{t+1} | I_t = 1, A_t] = B_t^\top \alpha + (A_t - \rho_t) Z_t^\top \beta \tag{2.2}$$

over all $t \in \{1, \ldots, T\}$, where ρ_t is the known randomization probability ($\Pr(A_t = 1) = \rho_t$) and the $q \times 1$ vector B_t is a function of t and covariates that are unaffected by treatment such as time of day or day of week. Note that A_t is centered by subtracting off the randomization probability; thus the working model for $\alpha(t) = \mathbb{E}[Y_{t+1} | I_t = 1]$ is $B_t^\top \alpha$. The estimators $\hat{\alpha},\ \hat{\beta}$ minimize the least squares error:

$$\mathbb{P}_N \left\{ \sum_{t=1}^{T} I_t \left(Y_{t+1} - B_t^\top \alpha - (A_t - \rho_t) Z_t^\top \beta \right)^2 \right\} \tag{2.3}$$

where $\mathbb{P}_N \big\{ f(X) \big\}$ is defined as the average of $f(X)$ over the sample.

Note that from a technical perspective, minimizing the least squares criterion, (2.3), is reminiscent of a GEE analysis [25] with identity link function and a working correlation matrix equal to the identity. Thus it is natural to consider a non-identity working correlation matrix as is common in GEE. This, however, is problematic from a causal inference perspective. To see this suppose that the true conditional expectation is in fact $\mathbb{E}\left[Y_{t+1} | I_t = 1, A_t\right] = B_t^\top \alpha^* + (A_t - \rho_t) Z_t^\top \beta^*$, that is, the causal parameter, $\beta(t)$ is equal to $Z_t^\top \beta^*$. Further suppose that the working correlation matrix has

off-diagonal elements and that we estimate β^* by minimizing the weighted (by the inverse of the working correlation matrix) least squares criterion. In this case the resulting estimating equations include sums of terms such as $I_t\left(Y_{t+1} - B_t^\top\alpha - (A_t-\rho_t)Z_t^\top\beta\right) I_s(A_s-\rho_t)Z_s$ for $t > s$. Unfortunately, both availability at time t, I_t, as well as Y_{t+1} may be affected by treatment in the past (in particular, A_s), thus absent strong assumptions $\mathbb{E}\left[I_t\left(Y_{t+1} - B_t^\top\alpha^* - (A_t-\rho_t)Z_t^\top\beta^*\right) I_s(A_s-\rho_t)\right]$ is unlikely to be 0. Recall that a minimal condition for consistency of estimators of (α^*, β^*) is that the estimating equations have expectation 0, thus absent further assumptions, the estimators derived from the weighted least squares criterion are likely biased. Another possibility is to include a time-varying variance term in the least squares criterion, that is the tth entry in (2.3) might be weighted by σ_t^{-2}. This would be useful in the data analysis, however for sample size calculations, values of these variances are unlikely to be available. Thus for simplicity we use the unweighted least squares criterion in (2.3).

Assume that the matrices $Q = \sum_{t=1}^T \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_tZ_t^\top$ and $\sum_{t=1}^T \mathbb{E}[I_t]B_tB_t^\top$ are invertible. The least squares estimators, $\hat\alpha$, $\hat\beta$ are consistent estimators of

$$\tilde\alpha = \left(\sum_{t=1}^T \mathbb{E}[I_t]B_tB_t^\top\right)^{-1} \sum_{t=1}^T \mathbb{E}[I_t]\alpha(t)B_t \tag{2.4}$$

and

$$\tilde\beta = \left(\sum_{t=1}^T \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_tZ_t^\top\right)^{-1} \sum_{t=1}^T \mathbb{E}[I_t]\rho_t(1-\rho_t)\beta(t)Z_t \tag{2.5}$$

respectively. Furthermore if $\beta(t)$ is in fact equal to $Z_t^\top\beta$ for some β, then $Z_t^\top\tilde\beta = \beta(t)$. This is the case even if $\mathbb{E}[Y_{t+1}|I_t=1] \neq B_t^\top\tilde\alpha$. In the Appendix (Lemma 1), we prove these results and also show that, under moment conditions, $\sqrt{N}(\hat\beta - \tilde\beta)$ is asymptotically normal with mean 0 and variance $\Sigma_\beta = Q^{-1}WQ^{-1}$ where,

$$W = \mathbb{E}\left[\left(\sum_{t=1}^T \tilde\epsilon_t I_t(A_t-\rho_t)Z_t\right) \times \left(\sum_{t=1}^T \tilde\epsilon_t I_t(A_t-\rho_t)Z_t^\top\right)\right]$$

and $\tilde\epsilon_t = Y_{t+1} - I_tB_t^\top\tilde\alpha - (A_t-\rho_t)I_tZ_t^\top\tilde\beta$. To test the null hypothesis $\mathrm{H}_0 : \beta(t) = 0,\ t = 1,\ldots,T$, one can use a test statistic based on the alternative, e.g.,

$$N\hat\beta^\top\hat\Sigma_\beta^{-1}\hat\beta \tag{2.6}$$

where $\hat\Sigma_\beta = \hat Q^{-1}\hat W\hat Q^{-1}$ and $\hat Q$ and $\hat W$ are plug in estimators. Note that this test statistic re-

sults from a GEE analysis with identity link function and a working correlation matrix equal to the identity matrix for which sample size formulae have been developed [26]. We build on this work as follows. As Tu *et al.* [26] discuss, under the null hypothesis the large sample distribution of this statistic is a chi-squared with p degrees of freedom distribution. If N, the sample size, is small, then, as recommended by Mancl and DeRouen [27], we make small adjustments to improve the small sample approximation to the distribution of the test statistic. In particular, they recommend adjusting $\hat{W}$ using the "hat" matrix; see the formulae for the adjusted $\hat{W}$ as well as $\hat{Q}$ in Appendix 2.8. Also in small sample settings, investigators commonly suggest that instead of using a critical value based on the chi-squared distribution, a critical value based on the $t-$distribution should be used [28]. As we are considering a simultaneous test for multiple parameters we form the critical value based on Hotelling's $T-$squared distribution [29]. Hotelling's $T-$squared distribution is a multiple of the F distribution given by $\frac{d_1(d_1+d_2-1)}{d_2}F_{d_1,d_2}$; here we use $d_1 = p$ and $d_2 = N - q - p$ (recall q is the number of parameters in the nuisance parameter vector, α); see the appendix for a rationale. In the following, the rejection region for the test of $\mathrm{H}_0 : \beta(t) = 0,\ t = 1, \ldots T$ based on (2.6) is

$$\left\{ N\hat{\beta}^\top \hat{\Sigma}_\beta^{-1}\hat{\beta} > \frac{p(N-q-1)}{N-q-p} F^{-1}_{p,N-q-p}\left(1-\alpha_0\right) \right\}$$

where α_0 is the desired significance level.

2.5 Sample Size Formulae

As Tu *et.al* [26] have developed general sample size formulas in the GEE setting, here we focus on considerations specific to the setting of micro-randomized trials. To size the study, we will determine the sample size needed to detect the alternate, $\beta(t)$ with:

$$\mathrm{H}_1 : \beta(t)/\bar{\sigma} = d(t),\ t = 1, \ldots, T$$

where $\bar{\sigma}^2 = (1/T)\sum_{t=1}^{T} \mathbb{E}\left[\mathrm{Var}\left(Y_{t+1}\middle| I_t = 1, A_t\right)\right]$ is the average variance and $d(t)$ is a standardized treatment effect. When N is large and H_1 holds, $N\hat{\beta}'\hat{\Sigma}_\beta^{-1}\hat{\beta}$ is approximately distributed as a noncentral chi-squared $\chi^2_p(c_N)$, where c_N, the non-centrality parameter, satisfies $c_N = N(\bar{\sigma}\tilde{d})'\Sigma_\beta^{-1}(\bar{\sigma}\tilde{d})$, and $\tilde{d} = \left(\sum_{t=1}^{T} \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t]\rho_t(1-\rho_t)d(t)Z_t$ [26]. Note that $\tilde{d} = \tilde{\beta}/\bar{\sigma}$.

Working Assumptions To derive the sample size formula, we use the form of the non-centrality parameter of the limiting non-central chi-squared distribution, along with working assumptions. The working assumptions are used to simplify the form of Σ_β^{-1}. In particular, we make the following working assumptions:

(a) $\mathbb{E}[Y_{t+1}|I_t = 1] = B_t^\top \alpha$, for some $\alpha \in \mathbb{R}^q$

(b) $\beta(t) = Z_t^\top \beta$ for some $\beta \in \mathbb{R}^p$

(c) $\text{Var}(Y_{t+1}|I_t = 1, A_t)$ is constant in t and A_t

(d) $\mathbb{E}[\tilde{\epsilon}_t \tilde{\epsilon}_s | I_t = 1, I_s = 1, A_t, A_s]$ is constant in A_t, A_s.

where, as before, $\tilde{\epsilon}_t = Y_{t+1} - I_t B_t^\top \tilde{\alpha} - (A_t - \rho_t) I_t Z_t^\top \tilde{\beta}$. See appendix 2.8 (Lemma 2) for proof of variance formulas under these working assumptions. The above working assumptions are somewhat simplistic but as will be seen below the resulting sample size formula is robust to moderate violations. First, under these working assumptions the alternative hypothesis can be re-written as

$$\text{H}_1 : \beta/\bar{\sigma} = d, \tag{2.7}$$

where d is a p dimensional vector of standardized effects. Furthermore, Σ_β is given by

$$\Sigma_\beta = \bar{\sigma}^2 \bigg(\sum_{t=1}^{T} \mathbb{E}[I_t] \rho_t (1 - \rho_t) Z_t Z_t^\top \bigg)^{-1},$$

and thus c_N is given by

$$c_N = N d^\top \bigg(\sum_{t=1}^{T} \mathbb{E}[I_t] \rho_t (1 - \rho_t) Z_t Z_t^\top \bigg) d. \tag{2.8}$$

To improve the small sample approximation, we use the multiple of the F-distribution as discussed above. Thus the sample size, N, is found by solving

$$F_{p,N-q-p;c_N} \left(F^{-1}_{p,N-q-p} \left(1 - \alpha_0\right) \right) = \beta_0 \tag{2.9}$$

where $F_{p,N-q-p;c_N}$ is the noncentral F distribution with noncentrality parameter, c_N and $1 - \beta_0$ is the desired power. The inputs to this sample size formula are $\{Z_t\}_{t=1}^T$, a scientifically meaningful value for d (see below for an illustration), the time-varying availability pattern, $\{\mathbb{E}[I_t]\}_{t=1}^T$, the desired significance level, α_0 and power, $1 - \beta_0$.

Table 2.1: Illustrative sample sizes for HeartSteps. The day of maximal treatment effect is 29. The expected availability is constant in t.

$\bar{d}$ \ $\mathbb{E}[I_t]$	0.7	0.6	0.5	0.4
0.10	32	36	42	52
0.09	38	44	51	63
0.08	47	54	64	78
0.07	60	69	81	101
0.06	79	92	109	135
0.05	112	130	155	193

$\bar{d} = (1/T)\sum_{t=1}^{T} Z_t^\top d$ is the average standardized treatment effect.

Now we describe how the information needed in the sample size formula might be obtained when the alternative is quadratic ($p = 3$, (2.1)). In this case we first elicit the initial standardized proximal main effect given by $Z_1^\top \beta/\bar{\sigma} = \beta_1/\bar{\sigma}$. Second we elicit the averaged across time, standardized proximal main effect $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top \beta/\bar{\sigma}$. Lastly we elicit the time at which the proximal main effect is maximal, i.e. $\arg\max_t Z_t^\top \beta$. These three quantities can then be used to solve for $d = (d_1, d_2, d_3)^\top$. For example, in HeartSteps, we might want to determine the sample size to ensure 0.80 power when there is no initial treatment effect on the first day, and the maximum proximal main effect comes around day 29. We specify the expected availability, $\mathbb{E}[I_t]$ to be constant in t and Z_t is given by (2.1). Table 2.1 gives sample sizes for HeartSteps under a variety of average standardized proximal main effects ($\bar{d}$).

In the behavioral sciences a standardized effect size of 0.2 is considered small [30]. Thus given the very small standardized effect sizes, the sample sizes given in Table 2.1 seem unbelievably small. Two points are worth making in this regard. First the use of the alternative parametric hypothesis (2.7) in forming the test statistic, implies that both between-subject as well as within-subject contrasts in proximal responses are used to detect the alternative. To see this, note that if we focused on only the first time point, $t = 1$, and tested $H_0 : \beta(1) = 0$, then an appropriate test would be a two-sample t-test based on the proximal response Y_2, in which case the required sample size would be much larger (akin to the sample size for a two arm randomized-controlled trial in which 40% of the subjects are randomized to the treatment arm). This two-sample t-test uses only between-subject contrasts in proximal response to test the hypothesis. The required sample size would be even larger for a test of $H_0 : \beta(1) = 0,\ \beta(2) = 0$ in which no relationship between $\beta(1)$

and $\beta(2)$ is assumed. Conversely the sample size would be smaller if one focused on detecting alternatives to $H_0 : \beta(1) = 0,\ \beta(2) = 0$ of the form $H_1 : \beta(1) = \beta(2) \neq 0$. The use of the alternative, $\beta(1) = \beta(2) \neq 0$, allows one to construct tests that use both between-subject as well as within-subject contrasts in proximal responses. Our approach is in between these two extremes in that we focus on detecting smooth, in t, alternatives to $H_0 : \beta(t) = 0$ for all t. This permits use of both within- as well as between-subject contrasts in proximal responses. The assumption of a parsimonious alternative enables the use of smaller sample sizes. A second point is that, at this time, there is no general understanding of how large the standardized effect size should be for these "in-the-moment" effects of a treatment. Thus these standardized effects may or may not be considered small in future.

2.6 Simulations

We consider a variety of simulations with different generative models to evaluate the performance of the sample size formulae. In the simulations presented here, we use the same setup as in HeartSteps; see Appendix 2.8 for simulations in other setups (Table 2.16 and 2.17). Specifically, the duration of the study is 42 days and there are 5 decision times within each day ($T = 210$). The randomization probability is 0.4, i.e. $\rho = \rho_t = P(A_t = 1) = 0.4$. The sample size formula is given in (2.8) and (2.9). All simulations are based on 1,000 simulated data sets.

Throughout this section the inputs to this sample size formula are $Z_t = \left(1, \lfloor \frac{t-1}{5} \rfloor, \lfloor \frac{t-1}{5} \rfloor^2\right)^\top$, the time-varying availability pattern, $\tau_t = \mathbb{E}[I_t]$, d, $\alpha_0 = .05$ and power, $1 - \beta_0 = .80$. The value for the vector d is indirectly specified via (a) the time at which the maximal standardized proximal main effect is achieved ($\arg\max_t Z_t^\top d$), (b) the averaged across time, standardized proximal main effect $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top d$ and (c) no initial standardized proximal main effect ($Z_1^\top d = d_1 = 0$). The test statistic used to evaluate the sample size formula is given by (2.6) in which B_t and Z_t are set to $\left(1, \lfloor \frac{t-1}{5} \rfloor, \lfloor \frac{t-1}{5} \rfloor^2\right)^\top$.

The simulation results provided below illustrate that the sample size formula and associated test statistic are robust. For convenience we summarize the results here. When the working assumptions hold, then under a variety of availability patterns, i.e., time-varying values for $\tau_t = \mathbb{E}[I_t]$ (see Figure 2.1) the desired Type I error and power are preserved. This is also the case when past treatment impacts availability. Furthermore the sample size formula is robust to deviations from the working assumptions, that is, provides the desired Type I error and power; this is true for a variety of forms of the true proximal main effect of the treatment (see Figure 2.2), a variety of distributions and correlation patterns for the errors, and dependence of Y_{t+1} on past treatment. In

all cases the above robustness occurs as long as we provide an approximately true or conservative value for the standardized effect, d and if we provide an approximately true or conservative (low) value for the availability, $\mathbb{E}[I_t]$.

In our simulations, we note several areas in which the sample size formula is less robust to the working assumption (c); this is when the error variance in Y_{t+1} varies depending on whether treatment $A_t = 1$ or $A_t = 0$ or with time t. In particular, if the ratio of conditional variance $\text{Var}[Y_{t+1}|I_t = 1, A_t = 1]/\text{Var}[Y_{t+1}|I_t = 1, A_t = 0] < 1$, then the power is reduced. Also if the average variance, $\mathbb{E}[\text{Var}[Y_{t+1}|I_t = 1, A_t]]$, varies greatly with time t, then the power is reduced. See below for details. Lastly as would be expected for any sample size formula, using values of the standardized effect size, d, or availability that are larger than the truth degrades the power of the procedure.

Working Assumptions Underlying Sample Size Formula are True

First, we considered a variety of settings in which the working assumptions (a)-(d) hold and in which the inputs to the sample size formula are correct (d is correct under the alternate hypothesis and the time-varying availability $\mathbb{E}[I_t]$ is correct). Neither the working assumptions nor the inputs to the sample size formula specify the error distribution, thus in the simulation we consider 5 distributions for the errors in the model for Y_{t+1} including independent normal, Student's t and exponential distributions as well as two autoregressive (AR) processes; all of these error patterns satisfy $\bar{\sigma}^2 = 1$ (recall $\bar{\sigma}^2 = (1/T)\sum_{t=1}^{T} \mathbb{E}\left[\text{Var}\left(Y_{t+1}\middle|I_t = 1, A_t\right)\right]$). Furthermore neither the working assumptions nor the inputs to the sample size formula specify the dependence of the availability indicator, I_t on past treatment. Thus we consider settings in which the availability decreases as the number of recent treatments increases. For brevity, we provide these standard results in the Appendix 2.8 (Tables 2.12, 2.13, 2.14 and 2.15). The results are generally quite good, with very few Type I error rates significantly above .05 and power levels significantly below .80.

Working Assumptions Underlying Sample Size Formula are False

Second, we considered a variety of settings in which the working assumptions are false but the inputs to the sample size formula are approximately correct as follows. Throughout $\bar{\sigma}^2 = 1$.

Working Assumption (a) is Violated. Suppose that the true $\mathbb{E}[Y_{t+1}|I_t = 1] \neq B_t\alpha$ for any $\alpha \in \mathbb{R}^q$. In particular, we consider the scenario in which there is a "weekend" effect on Y_{t+1}; see

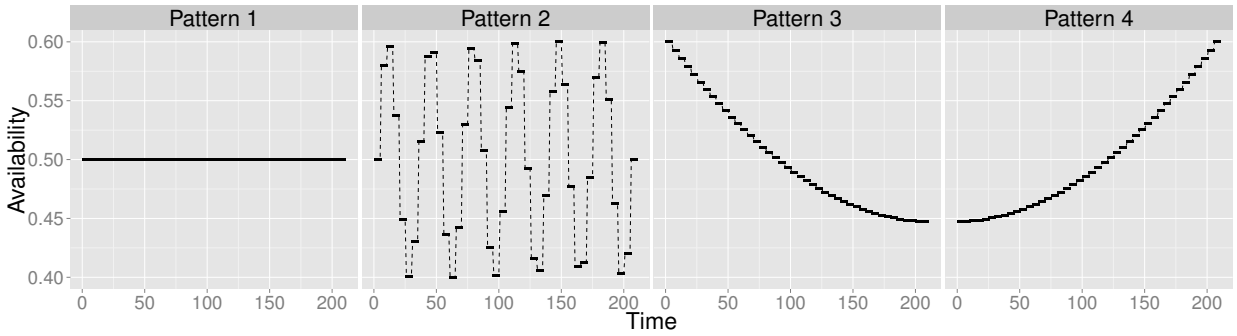

Figure 2.1: Availability Patterns. The x-axis is decision time point and y-axis is the expected availability. Pattern 2 represents availability varying by day of the week with higher availability on the weekends and lower mid-week. The average availability is 0.5 in all cases.

other scenario in Appendix 2.8. The data is generated as follows,

$$I_t \overset{Ber}{\sim} \big(\tau_t\big), \;\; A_t \overset{Ber}{\sim} \big(\rho\big)$$
$$Y_{t+1} = \alpha(t) + (A_t - \rho) Z_t^\top d + \epsilon_t, \text{ if } I_t = 1$$

where the conditional mean $\alpha(t) = B_t^\top \alpha + W_t \theta$. W_t is a binary variable: $W_t = 1$ if day of the week is time t is a weekend day, and $W_t = 0$ if the day is a weekday. For simplicity, we assume each subject starts on Monday, e.g., for $k = 1, \ldots, 6$, $W_{i+35(k-1)} = 0$, when $i = 1, \ldots, 25$, $W_{i+35(k-1)} = 1$, when $i = 26, \ldots, 35$ (recall that we assume in the simulation that there are 5 decision time points per day and the length of the study is 6 week). The values of $\{\alpha_i, i = 1, 2, 3\}$ are determined by setting $\alpha(1) = 2.5, \arg\max_t \alpha(t) = T, (1/T) \sum_{t=1}^{T} \alpha(t) - \alpha(1) = 0.1$. The error terms $\{\epsilon_t\}_{t=1}^{N}$ are i.i.d. $\mathrm{N}(0, 1)$. The day of maximal proximal effect is 29. Additionally, different values of the averaged standardized treatment effect and four patterns of availability as shown in Figure 2.1 with average 0.5 and are considered. The type I error rate is not affected, thus is omitted here. The simulated power is reported in Table 2.2; for more details see Table 2.19 in Appendix 2.8.

Working Assumption (b) is Violated. Suppose that the true $\beta(t) \neq Z_t^\top \beta$ for any β. Instead the vector of standardized effect, d, used in the sample size formula corresponds to the projection of $d(t)$, that is,

$$d = \left(\sum_{t=1}^{T} \mathbb{E}[I_t] Z_t Z_t^\top d \right)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t] Z_t d(t)$$

Table 2.2: Simulated power when working assumption (a) is violated. The patterns of availability are provided in Figure 2.1.

θ	$\bar{d}$	Availability Pattern		
		Pattern 1	Pattern 2	Pattern 3
$0.5\bar{d}$	0.10	0.80	0.79	0.81
	0.06	0.78	0.83	0.81
$1\bar{d}$	0.10	0.79	0.78	0.78
	0.06	0.78	0.79	0.79
$1.5\bar{d}$	0.10	0.78	0.81	0.78
	0.06	**0.77**	0.81	0.82
$2\bar{d}$	0.10	0.78	0.79	0.79
	0.06	0.81	0.79	0.78

θ is the coefficient of W_t in $\mathbb{E}[Y_{t+1}|I_t = 1]$. $\bar{d} = (1/T)\sum_{t=1}^{T} Z_t^\top d$ is the average standardized treatment effect. Bold numbers are significantly (at .05 level) greater lower than 0.80.

(recall $d(t) = \beta(t)/\bar{\sigma}$ and $\rho_t = \rho$). The sample size formula is used with the correct availability pattern, $\{\mathbb{E}[I_t]\}_{t=1}^{T}$. The data for each simulated subject is generated sequentially as follows. For each time t,

$$I_t \overset{Ber}{\sim} \big(\tau_t\big), \;\; A_t \overset{Ber}{\sim} \big(\rho\big)$$
$$Y_{t+1} = \alpha(t) + (A_t - \rho)d(t) + \epsilon_t, \text{ if } I_t = 1$$

for the variety of $d(t) = \beta(t)/\bar{\sigma}$ and $\mathbb{E}[I_t]$ patterns provided in Figure 2.2 and in Figure 2.1 respectively. The average availability is 0.5. The error terms $\{\epsilon_t\}_{t=1}^{T}$ are generated as i.i.d. $N(0,1)$. The conditional mean, $\mathbb{E}[Y_{t+1}|I_t = 1] = \alpha(t)$ is given by $\alpha(t) = \alpha_1 + \alpha_2\lfloor\frac{t-1}{5}\rfloor + \alpha_3\lfloor\frac{t-1}{5}\rfloor^2$, where $\alpha_1 = 2.5$, $\alpha_2 = 0.727$,$\alpha_3 = -8.66\times10^{-4}$ (so that $(1/T)\sum_t \alpha(t) - \alpha(1) = 1$, $\arg\max_t \alpha(t) = T$).

The simulated powers are provided in Table 2.3. In all cases the power is close to .80; this is because all of the proximal main effect patterns in Figure 2.2 are sufficiently well approximated by a quadratic in time. See Appendix 2.8 for other cases of $d(t)$ and details (Figure 2.5 and Table 2.9).

Working Assumption (c) is Violated. Suppose that $\text{Var}[Y_{t+1}|I_t = 1, A_t] = A_t\sigma_{1t}^2 + (1 - A_t)\sigma_{0t}^2$ where $\sigma_{1t}/\sigma_{0t} \neq 1$. The sample size formula is used with the correct pattern for $\{Z_t^\top d,\ \mathbb{E}[I_t]\}_{t=1}^{T}$.

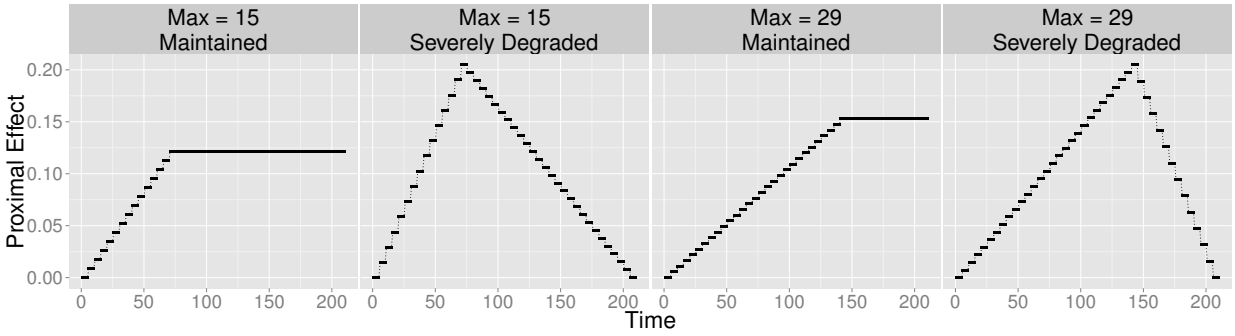

Figure 2.2: Standardized Proximal Main Effects of Treatment, $\{d(t)\}_{t=1}^{T}$: representing maintained and severely degraded time-varying proximal treatment effects. The horizontal axis is the decision time point. The vertical axis is the standardized treatment effect. The "Max" in the titles refer to the day of maximal proximal effect. The average standardized proximal effect is $\bar{d} = 0.1$ in all plots.

Table 2.3: Simulated power when working assumption (b) is violated. The shape of the standardized proximal effect and pattern for availability are provided in Figure 2.2 and 2.1 respectively. The sample sizes are given on the right.

$\bar{d}$	Availability Pattern	Max	Shape of $d(t)$		Sample Size	
			Maintained	Degraded		
0.10	Pattern 1	15	0.78	0.79	43	39
		29	0.80	0.79	38	38
	Pattern 2	15	0.79	0.80	43	39
		29	0.78	0.79	38	38
	Pattern 3	15	0.81	**0.77**	45	41
		29	0.81	0.78	37	39
0.06	Pattern 1	15	0.81	0.79	111	100
		29	0.81	0.79	96	96
	Pattern 2	15	0.79	0.81	112	100
		29	0.79	0.80	96	96
	Pattern 3	15	0.78	0.81	116	106
		29	0.80	0.80	95	101

$\bar{d} = (1/T)\sum_{t=1}^{T} Z_t^\top d$ is the average standardized treatment effect. The "Max" in the first row refers to the day of maximal proximal effect. Bold numbers are significantly (at .05 level) lower than .80.

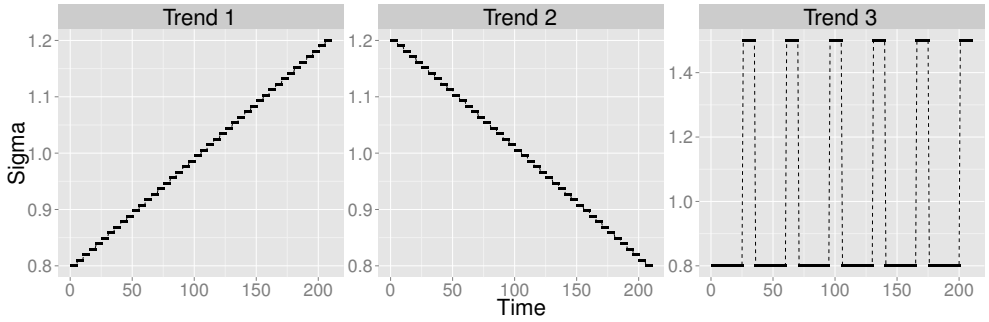

Figure 2.3: Trend of $\bar{\sigma}_t$: For all trends, $\bar{\sigma}_t^2$ is scaled so that $(1/T)\sum_{t=1}^{T}\bar{\sigma}_t^2 = 1$. In Trend 3, the variance, $\bar{\sigma}_t^2 = \mathbb{E}\big[Var[Y_{t+1}|I_t = 1, A_t]\big]$ peaks on weekends. In particular, $\bar{\sigma}_{7k+i} = 0.8$ for $i = 1, \ldots, 5$ and $\bar{\sigma}_{7k+i} = 1.5$ for $i = 6, 7$.

The data for each simulated subject is generated sequentially as follows. For each time t,

$$I_t \overset{Ber}{\sim} \big(\tau_t\big), \;\; A_t \overset{Ber}{\sim} \big(\rho\big)$$
$$Y_{t+1} = \alpha(t) + (A_t - \rho)Z_t^\top d + \mathbb{1}_{\{A_t=1\}}\sigma_{1t}\epsilon_t + \mathbb{1}_{\{A_t=0\}}\sigma_{0t}\epsilon_t, \text{ if } I_t = 1$$

where the average across time standardized proximal main effect, $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top d$ is 0.1 and day of maximal effect is equal to 22 or 29. The function $\alpha(t) = \mathbb{E}[Y_{t+1}|I_t = 1]$ is as in the prior simulation. The availability, $\tau_t = 0.5$. The error terms $\{\epsilon_t\}$ follow a normal AR(1) process, e.g., $\epsilon_t = \phi\epsilon_{t-1} + v_t$ with the variance of v_t scaled so that $\text{Var}[\epsilon_t] = 1$. Define $\bar{\sigma}_t^2 = \mathbb{E}\big[\text{Var}[Y_{t+1}|I_t = 1, A_t]\big]$ $(= \rho\sigma_{1t}^2 + (1-\rho)\sigma_{0t}^2)$. Recall the average variance $\bar{\sigma}^2$ is given by $(1/T)\sum_{t=1}^{T}\bar{\sigma}_t^2$. We consider 3 time-varying trends for $\{\bar{\sigma}_t\}$ together with different values of σ_{1t}/σ_{0t}; see Figure (2.3). In each trend, $\bar{\sigma}_t^2$ is scaled such that $\bar{\sigma} = 1$; thus the standardized proximal main effect in the generative model is $Z_t^\top d$. In all cases, the simulated type I error rates are close to $.05$ and thus the table is omitted here (see Appendix 2.8, Table 2.10). The simulated power is given in Table 2.4.

In the case of $\sigma_{1t} < \sigma_{0t}$, the simulated powers are slightly larger than 0.8, while the simulated powers are smaller than 0.8 in the case of $\sigma_{1t} > \sigma_{0t}$. The impact of $\bar{\sigma}_t$ on the power depends on the shape of treatment effect: when $\beta(t)$ attains its maximum, more than halfway through the study, at day 29, a increasing $\{\bar{\sigma}_t\}$, trend 1, lowers the power, while a decreasing $\{\bar{\sigma}_t\}$, trend 2, improves the power. When $\beta(t)$ attains a maximal effect midway through the study, either decreasing or increasing $\{\bar{\sigma}_t\}$ does not impact power. A large variation in $\bar{\sigma}_t$, e.g., trend 3, reduces the power in all cases. The differing auto correlations of the errors, ϵ_t, do not affect power; see a more detailed table in Appendix 2.8, Table 2.10.

Table 2.4: Simulated power when working assumption (c) is violated, $\sigma_{1t} \neq \sigma_{0t}$. The trends are provided in Figure 2.3. The availability is 0.5. The average proximal main effect, $\bar{d} = 0.1$ and the day of maximal effect is 22 or 29, and thus the associated sample sizes are 41 and 42.

ϕ	$\frac{\sigma_{1t}}{\sigma_{0t}}$	Max = 22 (N = 41)			Max = 29 (N = 42)		
		trend 1	trend 2	trend 3	trend 1	trend 2	trend 3
-0.6	0.8	0.83	0.84	0.80	0.81	0.89	0.79
	1.0	0.79	0.80	**0.75**	**0.74**	0.85	**0.70**
	1.2	**0.76**	**0.76**	**0.71**	**0.72**	0.81	**0.70**
0	0.8	0.85	0.82	0.79	0.81	0.88	0.78
	1.0	0.79	0.81	**0.74**	**0.77**	0.86	**0.72**
	1.2	**0.77**	**0.77**	**0.71**	**0.70**	0.83	**0.70**
0.6	0.8	0.83	0.83	0.81	**0.77**	0.87	**0.77**
	1.0	**0.76**	0.79	**0.75**	**0.73**	0.85	**0.77**
	1.2	0.78	**0.77**	**0.73**	**0.72**	0.82	**0.69**

ϕ is the parameter in AR(1) for $\{\epsilon_t\}_{t=1}^{T}$. "Max"is the day in which the maximal proximal effect is attained. Bold numbers are significantly (at .05 level) lower than .80.

Table 2.5: Simulated power when working assumption (d) is false. The expected availability is 0.5, the average proximal main effect $\bar{d} = 0.1$ and the maximal effect is attained at day 29. The associated sample size is 42.

Parameters in I_t	γ_1 \ γ_2	-0.1	-0.2	-0.3
$\eta_1 = -0.1, \eta_2 = -0.1$	-0.2	0.80	0.81	0.79
	-0.5	0.79	0.81	0.80
	-0.8	0.81	0.82	0.79
$\eta_1 = -0.2, \eta_2 = -0.1$	-0.2	0.78	0.82	0.79
	-0.5	0.81	**0.77**	**0.77**
	-0.8	0.81	0.79	0.78
$\eta_1 = -0.1, \eta_2 = -0.2$	-0.2	0.78	0.78	0.80
	-0.5	0.80	0.79	0.78
	-0.8	0.78	0.79	0.80

γ_1 and γ_2 are parameters for the cumulative treatments in the model of Y_{t+1}. η_1 and η_2 are parameters in the model of I_t. Bold numbers are significantly (at .05 level) less than .80.

Working Assumption (d) is Violated We violate assumption (d) by making both the availability indicator, I_t and proximal response, Y_{t+1} depend on past treatment and past proximal responses. The sample size formula is used with the correct value of $\{Z_t^\top d, \mathbb{E}[I_t]\}_{t=1}^T$; in particular d is determined by an average proximal main effect of $\bar{d} = 0.1$, day of maximal effect equal to 29 ($d_1 = 0, d_2 = 9.64 \times 10^{-3}, d_3 = -1.72 \times 10^{-4}$) and with a constant availability pattern equal to 0.5. The data for each simulated subject is generated as follows. Denote the cumulative treatment over last 24 hours by $C_t = \sum_{j=1}^5 A_{t-j} I_{t-j}$. In each time t,

$$I_t \overset{Ber}{\sim} \left(\tau_t + \tau_t \eta_1 (C_t - \mathbb{E}[C_t]) + \tau_t \eta_2 \,\mathrm{Trunc}(\frac{1}{5}\sum_{j=1}^5 \epsilon_{t-j})\right), \;\; A_t \overset{Ber}{\sim} \left(\rho\right)$$

$$\begin{aligned} Y_{t+1} = {} & \mathbb{1}_{\{I_t=0\}} \left(\alpha_0(t) + \epsilon_t\right) + \\ & \mathbb{1}_{\{I_t=1\}} \left(\alpha(t) + \gamma_1 \left[C_t - \mathbb{E}[C_t|I_t=1]\right] + (A_t - \rho)[Z_t^\top d + Z_t^\top \gamma_2 (C_t - \mathbb{E}[C_t|I_t=1])] + \sigma^* \epsilon_t\right) \end{aligned}$$

where $\{\epsilon_t\}_{t=1}^T$ are i.i.d $N(0,1)$ and $\mathrm{Trunc}(x) := x\mathbb{1}_{|x|\le 1} + \mathrm{sign}(x)\mathbb{I}_{|x|>1}$ (the truncation is used to ensure that $\tau_t + \tau_t\eta_1(C_t - \mathbb{E}[C_t]) + \tau_t\eta_2\,\mathrm{Trunc}(\frac{1}{5}\sum_{j=1}^5 \epsilon_{t-j}) \in [0,1]$). Again $\alpha(t)$ is as in the prior simulation. σ^* is calculated such that the average variance is equal to 1, e.g., $\bar{\sigma} = \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathrm{Var}[Y_{t+1}|I_t = 1, A_t]] = 1$. Note that since C_t is centered in both the model for I_t as well as in the model for Y_{t+1}, the standardized proximal main effect is $Z_t^\top d$ and $\mathbb{E}[I_t] = \tau_t = 0.5$. $\alpha_0(t)$ is the conditional mean of Y_{t+1} when $I_t = 0$. The form of $\mathbb{E}[Y_{t+1}|I_t = 0]$ is not essential: only $Y_{s+1} - \mathbb{E}[Y_{s+1}|I_s = 0]$ is used to generate I_t. In the simulation, $\mathbb{E}[C_t|I_t = 1]$ and σ^* are calculated by Monte Carlo methods. As before, the simulated type I error are not affected; see Table 2.7 in appendix 2.8. The simulated powers are provided in Table 2.5.

Some Practical Guidelines

Third, it is critical to use conservative values of d and availability $\mathbb{E}[I_t]$ in the sample size formula. It is not surprising that the quality of the sample size formula depends on an accurate or conservative values of the standardized effects, d, as this is the case for all sample size formulas. Additionally availability provides the number of decision points as which treatment might be provided per individual and thus the sample size formula should be sensitive to availability. To illustrate these points we consider two simulations in which the data is generated by

$$\begin{aligned} & I_t \overset{Ber}{\sim} \left(\tau_t\right), \;\; A_t \overset{Ber}{\sim} \left(\rho\right) \\ & Y_{t+1} = \alpha(t) + (A_t - \rho) Z_t^\top d + \epsilon_t, \text{ if } I_t = 1 \end{aligned}$$

where the ϵ_t's are i.i.d. standard normals and $\alpha(t)$ is as in the prior simulations. In the first simulation, suppose the scientist provides the correct availability pattern, $\{\mathbb{E}[I_t]\}_{t=1}^T$, the correct time at which the maximal standardized proximal main effect is achieved ($\arg\max_t Z_t^\top d$) and the correct initial standardized proximal main effect ($Z_1^\top d = d_1 = 0$) but provides too low a value of the averaged across time, standardized proximal main effect $\bar{d} = \frac{1}{T}\sum_{t=1}^T Z_t^\top d$. The simulated power is provided in Appendix 2.8, Table 2.20. The degradation in power is pronounced as might be expected.

In the second simulation, suppose the scientist provides the correct $\arg\max_t Z_t^\top d$, correct $Z_1^\top d = d_1 = 0$, correct $\bar{d} = \frac{1}{T}\sum_{t=1}^T Z_t^\top d$ and although the scientist's time-varying pattern of availability is correct, the magnitude is underestimated. The simulation result is in Appendix 2.8, Table 2.21. Again the degradation in power is pronounced.

2.7 Discussion

In this chapter, we have introduced the use of micro-randomized trials in mobile health and have provided an approach to determining the sample size. More sophisticated sample size procedures might be entertained. Certainly it makes sense to include baseline information in the sample size procedure, for example in HeartSteps, a natural baseline variable is baseline step count. The inclusion of baseline variables in B_t in the regression (2.2) is straightforward. An interesting generalization to the sample size procedure would allow scientists to include time-varying variables (in O_t) as covariates in B_t in the regression (2.2). This might be a useful strategy for reducing the error variance.

An alternate to the micro-randomized trial design is the single case design often used in the behavioral sciences [31]. These trials usually only involve 1 to 13 participants [32] and the data analyses focus on the examination of visual trends for each participant separately. For example, during periods when a participant is on treatment the response might be generally higher than the height of the response during the time periods in which the participant is off treatment. Dallery et al. [33] provide an excellent overview of single case designs and their use for evaluating technology based intervention. Their chapter illustrates the visual analyses that would be conducted on each participant's data. A critical assumption is that the effect of the treatment is only temporary (no carry-over effect) so that each participant can act as his own control. We believe that in settings in which treatments are expected to have sufficiently strong effects so as to overwhelm the within person variability in response (thus a visual analysis can be compelling), these designs provide an alternative to the micro-randomized trial design.

Although this chapter has focused on determining the sample size to detect the proximal main effect of a treatment with a given power, micro-randomized studies provide data for a variety of interesting further analyses. For example, it is of some interest to model and understand the predictors of the time-varying availability indicator. In the case of HeartSteps we will know why the participant is unavailable (driving a car, already active or has turned off the lock-screen messages) so we will be able to consider each type of availability indicator. Other very interesting further analyses include assessing interactions between treatments, A_t and context, O_t, past treatment $A_s, s < t$ on the proximal response, Y_{t+1}. Also there is much interest in using this type of data to construct "dynamic treatment regimes"; in this setting these are called Just-in-Time Adaptive Interventions [16]. The sequential micro-randomizations enhance all of these analyses by reducing causal confounding.

2.8 Appendix

A. Theoretical Results and Proofs

Lemma 1 (Least Squares Estimator)**.** *The least square estimators $\hat{\alpha}, \hat{\beta}$ are consistent estimators of $\tilde{\alpha}, \tilde{\beta}$ in* (2.4) *and* (2.5)*. In particular, if $\beta(t) = Z_t^\top \beta^*$ for some vector β^*, then $\tilde{\beta} = \beta^*$. Under moment conditions, we have $\sqrt{N}(\hat{\beta} - \tilde{\beta}) \to N(0, \Sigma_\beta)$, where the asymptotic variance Σ_β is given by $\Sigma_\beta = Q^{-1}WQ^{-1}$ where $Q = \sum_{t=1}^T \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_tZ_t^\top$, $W = \mathbb{E}\Big[\sum_{t=1}^T \tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t \times \sum_{t=1}^T \tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t^\top\Big]$ and $\tilde{\epsilon}_t = Y_{t+1} - B_t^\top \tilde{\alpha} - Z_t^\top \tilde{\beta}(A_t - \rho_t)$.*

Proof. It's easy to see that the least square estimators satisfy

$$\begin{aligned}\hat{\theta} = (\hat{\alpha}, \hat{\beta}) &= \big(\mathbb{P}_N \sum_{t=1}^T I_t X_t X_t^\top\big)^{-1}\big(\mathbb{P}_N \sum_{t=1}^T I_t Y_{t+1} X_t\big) \\ &\to \big(\sum_{t=1}^T \mathbb{E}[I_t X_t X_t^\top]\big)^{-1}\big(\sum_{t=1}^T E(I_t Y_{t+1} X_t)\big)\end{aligned}$$

where $X_t^\top = (B_t^\top, (A_t - \rho_t)Z_t^\top) \in \mathbb{R}^{1\times(p+q)}$ is the covariate at time t. For each t,

$$\mathbb{E}[I_t Y_{t+1} X_t] = \begin{pmatrix} \mathbb{E}[I_t Y_{t+1}]B_t \\ \mathbb{E}[I_t Y_{t+1}(A_t - \rho_t)]Z_t \end{pmatrix} = \begin{pmatrix} \mathbb{E}[I_t Y_{t+1}]B_t \\ \rho_t(1-\rho_t)\mathbb{E}[I_t]\beta(t)Z_t \end{pmatrix},$$

and

$$\mathbb{E}[I_t X_t X_t^\top] = \begin{pmatrix} \mathbb{E}[I_t] B_t B_t^\top & B_t Z_t^\top \mathbb{E}[I_t(A_t - \rho_t)] \\ Z_t B_t^\top \mathbb{E}[I_t(A_t - \rho_t)] & Z_t Z_t^\top \mathbb{E}[I_t(A_t - \rho_t)^2] \end{pmatrix}$$
$$= \begin{pmatrix} \mathbb{E}[I_t] B_t B_t^\top & 0 \\ 0 & \mathbb{E}[I_t] \rho_t(1-\rho_t) Z_t Z_t^\top \end{pmatrix}$$

so that

$$\hat{\alpha} \to \left(\sum_{t=1}^{T} \mathbb{E}[I_t] B_t B_t^\top \right)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t Y_{t+1}] B_t = \left(\sum_{t=1}^{T} \mathbb{E}[I_t] B_t B_t^\top \right)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t] \alpha(t) B_t$$

$$\hat{\beta} \to \left(\sum_{t=1}^{T} \rho_t(1-\rho_t) \mathbb{E}[I_t] Z_t Z_t^\top \right)^{-1} \sum_{t=1}^{T} E[I_t Y_{t+1}(A_t - \rho_t)] Z_t$$
$$= \left(\sum_{t=1}^{T} \rho_t(1-\rho_t) \mathbb{E}[I_t] Z_t Z_t^\top \right)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t] \rho_t(1-\rho_t) \beta(t) Z_t$$

as in (2.4) and (2.5). We can see that if $\beta(t) = Z_t^\top \beta^*$, then

$$\big(\sum_{t=1}^{T} \rho_t(1-\rho_t) \mathbb{E}[I_t] Z_t Z_t^\top \big)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t] \rho_t(1-\rho_t) \beta(t) Z_t$$
$$= \big(\sum_{t=1}^{T} \rho_t(1-\rho_t) \mathbb{E}[I_t] Z_t Z_t^\top \big)^{-1} \sum_{t=1}^{T} \mathbb{E}[I_t] \rho_t(1-\rho_t) Z_t Z_t^\top \beta^* = \beta^*$$

This is true even if $\mathbb{E}[Y_{t+1} | I_t = 1] \neq B_t^\top \tilde{\alpha}$. We can easily see that,

$$\sqrt{N}(\hat{\theta} - \tilde{\theta}) = \sqrt{N} \left\{ \big(\mathbb{P}_N \sum_{t=1}^{T} I_t X_t X_t^\top \big)^{-1} \Big[\big(\mathbb{P}_N \sum_{t=1}^{T} I_t Y_{t+1} X_t \big) - \big(\mathbb{P}_N \sum_{t=1}^{T} I_t X_t X_t^\top \big) \tilde{\theta} \Big] \right\}$$
$$= \sqrt{N} \left\{ \mathbb{E}\big[\sum_{t=1}^{T} I_t X_t X_t^\top \big]^{-1} \big(\mathbb{P}_N \sum_{t=1}^{T} I_t \tilde{\epsilon}_t X_t \big) \right\} + o_p(\mathbf{1}), \tag{2.10}$$

where $o_p(\mathbf{1})$ is a term that converges in probability to zero as N goes to infinity. By the definitions

of $\tilde{\alpha}$ and $\tilde{\beta}$, we have

$$\mathbb{E}\big[\sum_{t=1}^{T} I_t\tilde{\epsilon}_t X_t\big] = \begin{pmatrix} \sum_{t=1}^{T}\mathbb{E}[I_t]\left(\alpha(t) - B_t^\top\tilde{\alpha}\right)B_t \\ \sum_{t=1}^{T}\mathbb{E}[I_t]\rho_t(1-\rho_t)\big(\beta(t) - Z_t^\top\tilde{\beta}\big)Z_t \end{pmatrix} = \mathbf{0}$$

So that under moments conditions, we have $\sqrt{N}(\hat{\theta} - \tilde{\theta}) \to N(0, \Sigma_\theta)$, where Σ_θ is given by

$$\Sigma_\theta = \mathbb{E}\big[\sum_{t=1}^{T} I_t X_t X_t^\top\big]^{-1}\mathbb{E}\big[\sum_{t=1}^{T} I_t\tilde{\epsilon}_t X_t \times \sum_{t=1}^{T} I_t\tilde{\epsilon}_t X_t^\top\big]\mathbb{E}\big[\sum_{t=1}^{T} I_t X_t X_t^\top\big]^{-1} = \begin{bmatrix} \Sigma_\alpha & \Sigma_{\alpha\beta} \\ \Sigma_{\alpha\beta}^\top & \Sigma_\beta \end{bmatrix}.$$

In particular, $\hat{\beta}$ satisfies $\sqrt{N}(\hat{\beta} - \tilde{\beta}) \to N(0, \Sigma_\beta)$ and $\Sigma_\beta = Q^{-1}WQ^{-1}$ where

$$Q = \sum_{t=1}^{T}\mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top,\ W = \mathbb{E}\bigg[\sum_{t=1}^{T}\tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t \times \sum_{t=1}^{T}\tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t^\top\bigg]$$

□

Lemma 2 (Asymptotic Variance Under Working Assumptions)**.** *Assuming working assumptions (2.5)-(2.5) are true, then under the alternative hypothesis H_1 in (2.7), Σ_β and c_N are given by*

$$\Sigma_\beta = \bar{\sigma}^2\bigg(\sum_{t=1}^{T}\mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top\bigg)^{-1},$$

$$c_N = Nd^\top\bigg(\sum_{t=1}^{T}\mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top\bigg)d.$$

Proof. Note that under assumptions (2.5) and (2.5), we have $Z_t^\top\tilde{\beta} = \beta(t)$ and $\text{Var}(Y_{t+1}|I_t = 1, A_t) = \bar{\sigma}$ for each t, and $\tilde{d} = d$. The middle term, W, in Σ_β can be separated by two terms, e.g.,

$$\begin{aligned} W &= \mathbb{E}\big[\sum_{t=1}^{T}\tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t \times \sum_{t=1}^{T}\tilde{\epsilon}_t I_t(A_t - \rho_t)Z_t^\top\big] \\ &= \sum_{t=1}^{T}\mathbb{E}\big[\tilde{\epsilon}_t^2 I_t(A_t - \rho_t)^2\big]Z_t Z_t^\top + \sum_{i\neq j}^{T}\mathbb{E}\big[\tilde{\epsilon}_i\tilde{\epsilon}_j I_i I_j(A_i - \rho_i)(A_j - \rho_j)\big]Z_i Z_j^\top. \end{aligned}$$

Under assumptions (2.5), (2.5) and (2.5), we have $\mathbb{E}[\tilde{\epsilon}_t|I_t = 1, A_t] = 0$ and $\mathbb{E}\big[\tilde{\epsilon}_t^2 I_t(A_t - \rho_t)^2\big] = \mathbb{E}[I_t]\rho_t(1-\rho_t)\bar{\sigma}^2$. Furthermore, suppose $i > j$, then $\mathbb{E}\big[\tilde{\epsilon}_i\tilde{\epsilon}_j I_i I_j(A_i - \rho)(A_j - \rho)\big] = \mathbb{E}[I_i I_j(A_j - \rho)(A_i - \rho)] \times \mathbb{E}[\tilde{\epsilon}_t\tilde{\epsilon}_s|I_t = 1, I_s = 1, A_t, A_s] = 0$, because $A_i \perp\!\!\!\perp \{I_i, I_j, A_j\}$ and the first term is 0.

W is then given by

$$W = \bar{\sigma}^2 \sum_{t=1}^{T} \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top,$$

so that $\Sigma_\beta = \bar{\sigma}^2 \big(\sum_{t=1}^{T} \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top \big)^{-1}$ and

$$c_N = N(\bar{\sigma}\tilde{d})^\top \Sigma_\beta^{-1}(\bar{\sigma}\tilde{d}) = N d^\top \left(\sum_{t=1}^{T} \mathbb{E}[I_t]\rho_t(1-\rho_t)Z_t Z_t^\top \right) d$$

□

Remark: Working assumption (d) can be replaced by assuming $\mathbb{E}[Y_{t+1}|I_t = 1, A_t, I_s = 1, A_s] - \mathbb{E}[Y_{t+1}|I_t = 1, A_t]$ does not depend on A_t for any $s < t$, or some Markovian type of assumption, e.g., $Y_{t+1} \perp\!\!\!\perp \{Y_{s+1}, I_s, A_s, s < t\} | I_t, A_t$. Either of them implies $\mathbb{E}\big[\tilde{\epsilon}_i \tilde{\epsilon}_j I_i I_j (A_i - \rho_i)(A_j - \rho_j)\big] = 0$, so that Σ_β and c_N have the same simplified forms.

Rationale for multiple of F distribution The distribution of $n(\bar{X} - \mu)^\top \hat{\Sigma}^{-1}(\bar{X} - \mu)$ constructed from a random sample of size n of $N(\mu, \Sigma)$ random variables in which $\hat{\Sigma}$ is the sample covariance matrix follows a Hotelling's T-squared distribution. The Hotelling's T-squared distribution is a multiple of the F distribution, $\frac{d_1(d_1+d_2-1)}{d_2} F_{d_1,d_2}$ in which d_1 is the dimension of μ, and d_2 is the sample size. Our sample sample approximation replaces d_1 by p (the number of parameters in the test statistic) and d_2 by $n - q - p$ (sample size minus number of nuisance parameters minus d_1).

Formula for adjusted $\hat{W}$ and $\hat{Q}$ Define a individual-specific residual vector $\hat{e}$ as the $T \times 1$ vector with tth entry $\hat{e}_t = Y_{t+1} - I_t B_t^\top \hat{\alpha} - I_t(A_t - \rho_t) Z_t^\top \hat{\beta}$. For each individual define the tth row of the $T \times (p+q)$ individual-specific matrix X by $(I_t B_t^\top, I_t(A_t - \rho_t) Z_t)$. Then define $H = X \left[\mathbb{P}_N X^\top X \right]^{-1} X^\top$. The matrix $\hat{Q}^{-1}$ is the lower right $p \times p$ block in the inverse of $\mathbb{P}_N X^\top X$; the matrix $\hat{W}$ is the lower right $p \times p$ block in $\mathbb{P}_N \left[X^T (I - H)^{-1} \hat{e} \hat{e}^\top (I - H)^{-1} X \right]$.

B. Further Simulations and Details

B1. Simulation Results When Working Assumptions are True

We conduct a variety of simulations in settings in which the working assumptions hold, the scientist provides the correct pattern for the expected availability, $\tau_t = \mathbb{E}[I_t]$ and under the alternate, the standardized proximal main effect is $d(t) = Z_t^\top d$. Here we will mainly focus on the setup where

the duration of the study is 42 days and there are 5 decision times within each day, but similar results can be obtained in different setups; see below. The randomization probability is 0.4, i.e. $\rho = \rho_t = P(A_t = 1) = 0.4$. The sample size formula is given in (2.8) and (2.9). The test statistic is given by (2.6) in which B_t and Z_t equal to $\left(1, \lfloor \frac{t-1}{5} \rfloor, \lfloor \frac{t-1}{5} \rfloor^2\right)^\top$. All simulations are based on 1,000 simulated data sets. The significance level is 0.05 and the desired power is 80%.

In the first simulation, the data for each simulated subject is generated sequentially as follows. For $t = 1, \ldots, T = 210$, I_t, A_t and Y_{t+1} are generated by

$$
\begin{aligned}
&I_t \overset{Ber}{\sim} \left(\tau_t\right), \;\; A_t \overset{Ber}{\sim} \left(\rho\right) \\
&Y_{t+1} = \alpha(t) + (A_t - \rho) d(t) + \epsilon_t, \text{ if } I_t = 1
\end{aligned}
$$

where $d(t) = Z_t^\top d$ and τ_t are same as in the sample size model. The conditional mean, $\mathbb{E}[Y_{t+1}|I_t = 1] = \alpha(t)$ is given by $\alpha(t) = \alpha_1 + \alpha_2 \lfloor \frac{t-1}{5} \rfloor + \alpha_3 \lfloor \frac{t-1}{5} \rfloor^2$, where $\alpha_1 = 2.5$, $\alpha_2 = 0.727$,$\alpha_3 = -8.66 \times 10^{-4}$ (so that $(1/T)\sum_t \alpha(t) - \alpha(1) = 1$, $\arg\max_t \alpha(t) = T$). We consider 5 differing distributions for the errors $\{\epsilon_t\}_{t=1}^T$: independent normal; independent (scaled) Student's t distribution with 3 degrees of freedom; independent (centered) exponential distribution with $\lambda = 1$; a Gaussian AR(1) process, e.g., $\epsilon_t = \phi \epsilon_{t-1} + v_t$, where v_t is white noise with variance σ_v^2 such that Var$(\epsilon_t) = 1$; and lastly a Gaussian AR(5) process, e.g., $\epsilon_t = \frac{\phi}{5} \sum_{j=1}^5 \epsilon_{t-j} + v_t$, where v_t is white noise with variance σ_v^2 such that Var$(\epsilon_t) = 1$. In all cases the errors are scaled to have mean 0 and variance 1 (i.e. $\mathbb{E}[\epsilon_t | I_t = 1] = 0$, Var$[\epsilon_t | A_t, I_t = 1] = 1$). Additionally four availability patterns, e.g., time varying values for $\tau_t = \mathbb{E}[I_t]$, are considered; see Figure (2.1). The simulated type 1 error rate and power when the duration of study is 42 days are reported in Table 2.12, 2.13, 2.14 and 2.15. The simulation results in other setups, e.g., the length of the study is 4 week and 8 week, are reported in Table 2.16 and 2.17. The associated sample sizes are given in Table 2.11.

Since neither the working assumptions nor the inputs to the sample size formula specify the dependence of the availability indicator, I_t on past treatment. In the second simulation, we consider the setting in which the availability decreases as the number of treatments provided in the recent past increase. In particular, the data are generated as follows,

$$
\begin{aligned}
&I_t \overset{Ber}{\sim} \left(\tau_t + \eta \sum_{j=1}^5 (A_{t-j} I_{t-j} - \mathbb{E}[A_{t-j} I_{t-j}])\right), \;\; A_t \overset{Ber}{\sim} \left(\rho\right) \\
&Y_{t+1} = \alpha(t) + (A_t - \rho) d(t) + \epsilon_t, \text{ if } I_t = 1
\end{aligned}
$$

Note that since we center $\sum_{j=1}^5 A_{t-j} I_{t-j}$ in the generative model of I_t, the expected availability is τ_t. The specification of $\alpha(t)$, $\beta(t)$ and ϵ_t are same as in the first simulation. The simulated type I

error rate and power are reported Table 2.18.

B2. Further Details When Working Assumptions are False

Working Assumption (a) is Violated. Here we consider another setting in which the working assumption (a) is violated, e.g., the underlying true $\mathbb{E}[Y_{t+1}|I_t = 1]$ follows a non-quadratic form (recall that B_t is given by $\left(1, \lfloor\frac{t-1}{5}\rfloor, \lfloor\frac{t-1}{5}\rfloor^2\right)^\top$). The data is generated as follows

$$I_t \overset{Ber}{\sim} \left(\tau_t\right), \;\; A_t \overset{Ber}{\sim} \left(\rho\right)$$
$$Y_{t+1} = \alpha(t) + (A_t - \rho)Z_t^\top d + \epsilon_t, \text{ if } I_t = 1$$

where $\alpha(t) = \mathbb{E}[Y_{t+1}|I_t = 1]$ is provided in Figure 2.4. For each case, $\alpha(t)$ satisfies $\alpha(1) = 2.5$ and $(1/T)\sum_{t=1}^{T} -\alpha(1) = 0.1$. The error terms $\{\epsilon_t\}_{t=1}^{N}$ are i.i.d N$(0, 1)$. The day of maximal proximal effect is assumed to be 29. Additionally, different values of averaged standardized treatment effect and four patterns of availability in Figure 2.1 with average 0.5 are considered. The simulation results are reported in Table 2.6.

Additional Simulation Results When Other Working Assumptions are False The main body of the chapter reports part of the results when working assumptions (b), (c) and (d) are violated. Additional simulation results are provided here. In particular, the simulation result is reported in Table 2.9 when $d(t)$ follows other non-quadratic forms, e.g., working assumption (b) is false; see Figure 2.5. The simulated Type I error rate and power when working assumption (c) is false are reported in Table 2.10. The simulated Type I error rate when working assumption (d) is violated is reported in Table 2.7.

Simulation Results when $\bar{d}$ and $\bar{\tau}$ are misspecified. As discussed in the chapter, the first scenario considers the setting in which the scientist provides the correct availability pattern, $\{\mathbb{E}[I_t]\}_{t=1}^{T}$, the correct time at which the maximal standardized proximal main effect is achieved, $\arg\max_t Z_t^\top d$, and the correct initial standardized proximal main effect, $Z_1^\top d = d_1 = 0$, but provides too low a value of the averaged across time, standardized proximal main effect $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top d$. The simulated power is provided in Table 2.20. In the second scenario, the scientist provides the correct $\arg\max_t Z_t^\top d$, correct $Z_1^\top d = d_1 = 0$, correct $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top d$ and although the scientist's time-varying pattern of availability is correct, the magnitude, e.g., the average availability, is underestimated. The simulation result is in Table 2.21.

Table 2.6: Simulated Type I error rate (%) and power (%) when working assumption (a) is violated. Scenario 2. The shapes of $\alpha(t) = \mathbb{E}[Y_{t+1}|I_t = 1]$ and patterns of availability are provided in Figure 2.4 and Figure 2.1. The average availability is 0.5. The day of maximal proximal effect is 29. The associated sample size is given in Table 2.11.

		Availability Pattern							
$\alpha(t)$	$\bar{d}$	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 1	Pattern 2	Pattern 3	Pattern 4
	0.10	3.6	4.3	4.7	4.5	**77.4**	80.2	**76.2**	**75.9**
Shape 1	0.08	5.9	3.8	4.1	3.4	79.7	80.1	78.9	80.6
	0.06	4.6	5.7	4.2	**6.5**	78.7	**76.3**	78.3	79.9
	0.10	4.8	4.8	4.4	4.1	79.2	79.1	78.5	79.7
Shape 2	0.08	3.9	5.4	4.8	4.3	77.7	80.4	**76.8**	80.9
	0.06	5.1	5.5	3.4	4.9	78.3	79.4	79.8	80.2
	0.10	5.1	3.5	4.3	4.4	79.1	79.4	**75.6**	78.0
Shape 3	0.08	4.6	5.0	6.2	3.8	78.3	78.1	79.1	78.1
	0.06	4.8	4.4	5.4	4.2	78.0	78.3	79.8	77.7

$\bar{d} = (1/T)\sum_{t=1}^{T} Z_t^\top d$ is the average standardized treatment effect. Bold numbers are significantly (at .05 level) greater than .05 (for type I error rate) and lower than 0.80 (for power).

Table 2.7: Simulated Type I error rate (%) when working assumption (d) is violated. $\mathbb{E}[I_t] = 0.5$. The proximal effect $Z_t^\top d$ satisfies the average is 0.1 and day of maximal effect is 29. N = 42.

Parameters in I_t	γ_1 \ γ_2	-0.1	-0.2	-0.3
$\eta_1 = -0.1, \eta_2 = -0.1$	-0.2	5.7	3.2	3.9
	-0.5	3.2	4.2	4.9
	-0.8	4.2	5.1	5.5
$\eta_1 = -0.2, \eta_2 = -0.1$	-0.2	5.4	3.8	3.9
	-0.5	4.4	4.4	4.8
	-0.8	4.7	4.3	4.6
$\eta_1 = -0.1, \eta_2 = -0.2$	-0.2	4.5	5.0	5.0
	-0.5	4.9	3.8	6.0
	-0.8	4.7	4.8	4.8

η_1, η_2 are parameters in generating I_t. γ_1, γ_2 are coefficients in the model of Y_{t+1}. All numbers in this table are significantly (at .05 level) greater than .05.

Table 2.8: Sample Sizes when working assumption (b) is violated. The vector of standardized effects sizes, d, used in the sample size formula provides the projection of $d(t)$. The sample size formula is used with the correct availability pattern, $\{\mathbb{E}[I_t]\}_{t=1}^{T}$. The shape of the standardized proximal effect $d(t)$ and pattern for availability $\mathbb{E}[I_t]$ are provided in Figure 2.5 and in Figure (2.1). The significance level and desired power is 0.05 and 0.80.

$\bar{d}$	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$ = 0.7		
			Shape of $d(t)$					
			Maintained	Slightly Degraded	Severely Degraded	Maintained	Slightly Degraded	Severely Degraded
0.10	Pattern 1	15	43	41	39	32	31	29
		22	43	41	40	33	31	30
		29	38	37	38	29	28	29
	Pattern 2	15	43	41	39	33	31	30
		22	43	42	40	33	31	30
		29	38	37	38	29	28	29
	Pattern 3	15	45	43	41	33	32	31
		22	44	43	42	33	32	31
		29	37	38	39	28	28	29
	Pattern 4	15	42	39	37	32	30	28
		22	44	41	39	33	31	30
		29	39	38	38	29	28	28
0.08	Pattern 1	15	65	61	58	48	45	43
		22	65	62	60	48	46	44
		29	56	55	56	42	41	42
	Pattern 2	15	65	61	59	48	45	43
		22	65	62	60	48	46	44
		29	56	55	56	42	41	42
	Pattern 3	15	67	64	62	49	47	45
		22	66	64	63	48	47	46
		29	56	56	59	41	41	43
	Pattern 4	15	63	59	55	47	44	41
		22	65	61	58	48	45	43
		29	58	56	56	43	41	41
0.06	Pattern 1	15	111	105	100	81	76	73
		22	112	106	103	81	77	75
		29	96	94	96	70	69	70
	Pattern 2	15	112	105	100	81	77	73
		22	112	106	103	81	77	75
		29	96	94	96	70	68	70
	Pattern 3	15	116	111	106	83	79	76
		22	114	110	108	82	79	78
		29	95	96	101	69	69	72
	Pattern 4	15	108	100	94	79	74	70
		22	112	105	99	81	76	73
		29	100	95	95	72	69	70

Table 2.9: Simulated power (%) when working assumption (b) is violated. The shape of the standardized proximal effect, $d(t) = \beta(t)/\bar{\sigma}$ and pattern for availability, $\mathbb{E}[I_t]$ are provided in Figure 2.5 and in Figure (2.1). The corresponding sample sizes are given in Table 2.8.

$\bar{d}$	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$ = 0.7		
			Shape of $d(t)$					
			Maintained	Slightly Degraded	Severely Degraded	Maintained	Slightly Degraded	Severely Degraded
0.10	Pattern 1	15	78.4	78.8	78.6	79.1	80.1	77.6
		22	80.4	79.5	81.2	80.0	**76.9**	77.9
		29	80.4	79.2	78.9	**77.3**	**76.8**	81.1
	Pattern 2	15	78.6	79.9	79.9	80.1	80.4	81.3
		22	78.3	81.2	78.8	79.2	80.8	80.5
		29	77.9	80.8	79.3	78.1	77.7	82.2
	Pattern 3	15	81.0	79.7	**77.4**	77.9	80.9	77.6
		22	78.9	79.1	80.0	79.7	79.4	**75.9**
		29	80.9	77.5	77.7	80.6	79.2	78.5
	Pattern 4	15	79.7	79.5	77.9	79.5	81.7	78.0
		22	78.9	77.9	80.4	82.2	78.9	78.8
		29	77.9	79.7	79.0	78.0	80.2	80.8
0.08	Pattern 1	15	80.5	79.5	78.6	80.6	79.2	78.7
		22	78.9	78.7	78.8	78.9	80.7	80.3
		29	**76.6**	78.0	78.3	80.9	78.6	80.4
	Pattern 2	15	81.0	79.3	78.7	82.0	80.5	80.1
		22	82.4	80.6	80.0	78.0	79.6	79.4
		29	79.2	**76.9**	81.9	78.3	78.8	79.7
	Pattern 3	15	78.2	81.6	80.9	79.1	79.2	77.5
		22	80.9	79.5	78.6	79.2	78.3	81.4
		29	80.4	79.3	77.5	77.9	80.2	82.3
	Pattern 4	15	79.4	79.4	78.1	78.6	**77.4**	78.8
		22	81.3	78.4	78.4	80.6	79.4	80.4
		29	79.9	79.3	79.8	79.5	79.7	81.2
0.06	Pattern 1	15	81.2	80.5	79.0	77.8	78.7	79.6
		22	80.0	81.7	79.8	80.7	80.5	80.2
		29	81.2	78.7	79.2	81.2	79.7	80.1
	Pattern 2	15	78.7	77.5	81.4	80.7	81.0	80.7
		22	80.6	81.8	79.2	80.3	81.6	80.2
		29	78.5	80.2	80.0	77.7	78.1	78.0
	Pattern 3	15	78.1	80.0	80.9	79.7	79.3	78.8
		22	81.2	80.2	80.0	78.3	82.2	81.1
		29	79.6	81.6	79.8	80.2	81.6	**76.9**
	Pattern 4	15	78.2	79.8	78.9	79.5	**77.3**	79.2
		22	79.2	81.1	79.4	**76.8**	79.2	80.4
		29	79.9	78.5	79.8	80.1	78.9	81.8

Table 2.10: Simulated Type I error rate (%) and power (%) when working assumption (c) is violated. The trends of $\bar{\sigma}_t$ are provided in Figure 2.3. The standardized average effect is 0.1. $\mathbb{E}[I_t] = 0.5$. The associated sample sizes are 41, 42 when the day of maximal effect is 22, 29.

		Max = 22				Max = 29			
ϕ in AR(1)	$\frac{\sigma_{1t}}{\sigma_{0t}}$	const.	trend 1	trend 2	trend 3	const.	trend 1	trend 2	trend 3
	0.8	4.1	4.3	3.3	5.4	4.7	4.9	2.8	4.1
-0.6	1.0	4.6	5.0	4.0	4.4	4.4	4.8	4.2	4.3
	1.2	3.8	4.5	5.2	5.5	4.3	4.1	4.5	3.8
	0.8	5.2	4.7	4.0	3.4	5.4	4.9	6.2	4.5
-0.3	1.0	4.9	4.5	4.5	4.3	5.2	5.1	4.0	3.7
	1.2	5.4	4.6	4.1	3.8	3.7	5.2	4.3	5.0
	0.8	4.8	4.0	4.1	3.9	4.7	5.2	3.7	4.2
0	1.0	5.4	4.0	5.8	3.9	4.1	4.0	5.9	5.7
	1.2	4.4	4.9	5.0	4.6	3.7	4.8	4.4	4.9
	0.8	5.3	4.4	4.7	3.2	4.6	5.4	5.6	4.1
0.3	1.0	5.5	4.0	3.4	3.7	5.0	4.6	4.0	3.6
	1.2	3.8	4.5	4.5	4.8	4.5	5.0	6.2	4.3
	0.8	5.5	3.9	5.3	3.8	3.3	3.5	5.1	4.2
0.6	1.0	4.0	3.7	5.2	5.1	4.8	5.1	5.0	4.7
	1.2	4.5	5.1	4.6	4.9	4.5	4.4	4.7	4.8
	0.8	82.8	82.7	83.7	79.9	83.6	80.6	88.7	79.2
-0.6	1.0	81.1	79.1	79.9	**74.8**	77.7	**74.3**	84.8	**70.4**
	1.2	**76.6**	**76.3**	**76.3**	**70.6**	77.6	**72.0**	80.7	**70.4**
	0.8	83.0	83.0	86.0	80.3	82.7	79.2	87.9	78.0
-0.3	1.0	77.6	81.4	80.7	**74.9**	79.1	**74.5**	86.0	**73.7**
	1.2	78.2	**76.9**	**77.3**	**73.4**	**74.4**	**71.2**	81.0	**70.7**
	0.8	84.6	84.6	82.1	79.0	81.8	81.5	88.0	78.0
0	1.0	80.1	78.6	80.9	**73.6**	77.7	**76.5**	86.1	**71.8**
	1.2	**76.0**	**76.7**	**77.4**	**70.6**	**74.5**	**69.9**	83.4	**69.6**
	0.8	83.6	79.7	84.6	79.7	82.1	81.7	88.2	**75.7**
0.3	1.0	81.5	82.4	82.3	**73.9**	79.5	**74.6**	85.1	**71.5**
	1.2	**74.8**	**76.6**	78.2	**71.1**	**75.5**	**71.1**	82.5	**70.1**
	0.8	81.4	83.1	83.5	80.5	83.1	**77.1**	86.6	**76.9**
0.6	1.0	80.7	**76.4**	79.0	**74.8**	80.4	**73.4**	84.7	**76.8**
	1.2	**77.0**	77.5	**77.0**	**73.5**	**74.4**	**72.5**	81.6	**69.4**

ϕ is the parameter in AR(1) process for $\{\epsilon_t\}_{t=1}^T$. Bold numbers are significantly(at .05 level) greater than .05 (for type I error)and less than 0.80 (for power).

Table 2.11: Sample Sizes when the proximal treatment effect satisfies $d(t) = Z_t^\top d$. The significance level is 0.05. The desired power is 0.80.

Duration of Study	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
4-week	Pattern 1	15	59	89	154	43	65	112
		22	60	91	158	44	66	114
		29	58	87	152	43	64	110
	Pattern 2	15	59	89	154	43	65	112
		22	60	92	159	44	67	115
		29	58	89	154	43	64	111
	Pattern 3	15	59	90	157	44	66	113
		22	63	96	167	46	69	119
		29	62	94	163	45	67	115
	Pattern 4	15	59	89	155	43	65	112
		22	57	86	150	43	64	110
		29	54	82	142	41	61	105
6-week	Pattern 1	22	41	61	105	31	45	76
		29	42	64	109	32	47	79
		36	41	62	106	31	45	77
	Pattern 2	22	41	61	105	31	45	76
		29	43	64	110	32	47	80
		36	42	62	107	31	46	77
	Pattern 3	22	42	62	106	31	46	77
		29	44	66	114	33	48	82
		36	43	65	112	32	47	80
	Pattern 4	22	41	62	106	31	45	77
		29	41	62	106	31	46	78
		36	40	59	101	30	44	74
8-week	Pattern 1	29	32	47	80	25	35	58
		36	33	49	84	26	37	61
		43	33	48	82	25	36	60
	Pattern 2	29	32	47	80	25	35	58
		36	34	49	84	26	37	61
		43	33	49	82	25	36	60
	Pattern 3	29	33	48	82	25	36	59
		36	35	51	87	26	38	63
		43	34	50	86	26	37	62
	Pattern 4	29	33	48	81	25	36	59
		36	33	49	83	25	36	61
		43	32	47	80	25	35	59

Table 2.12: Simulated Type I error rate (%) when working assumptions are true. Duration of the study is 6-week. The associated sample size is given in Table 2.11.

Error Term	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
i.i.d. Normal	Pattern 1	22	3.8	4.5	4.9	4.6	5.3	4.8
		29	4.7	6.0	4.6	4.0	3.2	5.0
		36	5.0	5.4	4.9	4.3	4.8	4.6
	Pattern 2	22	4.8	4.1	4.8	4.4	3.5	4.1
		29	4.3	6.2	3.2	4.6	4.2	4.2
		36	4.5	4.8	5.2	4.5	3.5	5.4
	Pattern 3	22	4.7	4.5	6.3	4.4	4.9	4.9
		29	4.1	5.1	4.6	4.3	6.0	5.6
		36	4.7	4.4	4.6	4.1	5.1	4.4
	Pattern 4	22	5.4	3.5	4.5	4.8	4.7	5.0
		29	5.2	4.5	4.5	5.0	5.0	5.1
		36	3.8	4.1	5.4	4.7	5.0	5.9
i.i.d. t dist.	Pattern 1	22	4.3	4.4	3.2	4.1	4.1	5.2
		29	5.0	3.8	3.2	3.7	4.2	6.3
		36	4.3	4.5	4.0	5.0	5.7	5.4
i.i.d. Exp.	Pattern 1	22	4.5	4.6	4.4	3.7	**7.1**	3.1
		29	4.5	4.6	4.2	4.5	4.5	4.7
		36	2.7	4.8	4.8	3.9	3.7	3.4

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. ϕ is the parameter for AR(1) and AR(5) process. Bold numbers are significantly(at .05 level) greater than .05.

Table 2.13: Simulated Type I error rate (%) when working assumptions are true. Duration of the study is 6-week. The associated sample size is given in Table 2.11.

Error Term	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
AR(1) $\phi = -0.6$	Pattern 1	22	4.3	5.3	4.6	3.8	4.2	4.0
		29	4.6	5.4	5.1	4.0	4.4	4.3
		36	4.7	4.0	4.0	4.1	4.2	3.9
AR(1) $\phi = -0.3$	Pattern 1	22	5.8	3.4	4.4	3.3	4.0	5.4
		29	4.9	4.7	4.6	5.5	5.5	4.5
		36	4.0	4.7	4.4	4.9	5.0	4.7
AR(1) $\phi = 0.3$	Pattern 1	22	4.6	4.6	4.9	4.3	5.4	4.1
		29	4.8	5.3	4.1	4.3	4.2	5.2
		36	3.6	3.9	4.9	4.8	4.9	4.9
AR(1) $\phi = 0.6$	Pattern 1	22	4.4	5.1	4.9	3.6	5.2	3.7
		29	3.7	4.9	4.6	4.5	4.3	5.8
		36	4.4	**6.7**	5.2	5.6	3.6	5.1
AR(5) $\phi = -0.6$	Pattern 1	22	4.4	4.7	5.1	4.2	4.5	5.5
		29	4.3	5.1	4.3	3.2	3.5	4.2
		36	5.3	4.5	6.1	4.2	4.6	5.4
AR(5) $\phi = -0.3$	Pattern 1	22	3.7	4.4	6.0	5.0	4.5	3.5
		29	4.4	4.7	5.2	5.3	4.5	5.0
		36	4.5	5.0	5.1	4.1	5.3	4.8
AR(5) $\phi = 0.3$	Pattern 1	22	5.3	4.3	5.7	4.8	4.1	4.3
		29	3.9	4.8	4.1	4.0	4.3	4.9
		36	4.2	5.5	5.1	3.6	4.5	3.6
AR(5) $\phi = 0.6$	Pattern 1	22	5.1	4.5	4.0	4.5	3.8	5.2
		29	5.2	4.8	4.5	2.9	5.3	4.4
		36	4.1	3.6	4.6	3.9	4.4	4.9

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. ϕ is the parameter for AR(1) and AR(5) process. Bold numbers are significantly(at .05 level) greater than .05.

Table 2.14: Simulated power (%) when working assumptions are true. Duration of the study is 6-week. The associated sample size is given in Table 2.11

Error Term	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
i.i.d. Normal	Pattern 1	22	80.9	80.0	81.0	78.7	77.5	80.7
		29	78.4	80.6	77.8	80.6	78.7	79.0
		36	80.2	80.0	79.6	79.4	80.2	**77.0**
	Pattern 2	22	80.3	78.1	78.8	80.6	79.6	79.8
		29	80.3	79.1	80.2	**77.4**	79.9	79.9
		36	**76.8**	79.3	80.2	78.5	78.4	80.0
	Pattern 3	22	83.5	81.5	77.7	78.5	81.3	78.7
		29	77.9	79.1	78.5	77.8	78.8	79.0
		36	**77.3**	78.1	79.8	79.8	79.9	79.1
	Pattern 4	22	**77.2**	79.7	81.8	80.2	79.0	78.8
		29	80.1	78.8	80.3	79.4	80.6	80.1
		36	80.5	79.4	80.0	78.9	79.9	78.1
i.i.d. t dist.	Pattern 1	22	80.4	81.9	81.0	79.7	79.4	80.7
		29	81.7	82.2	82.2	79.1	82.3	**77.3**
		36	80.8	78.8	79.5	81.8	81.6	79.9
i.i.d. Exp.	Pattern 1	22	81.0	81.6	79.7	**77.2**	80.1	80.2
		29	80.6	82.4	80.3	79.0	79.8	80.3
		36	82.1	79.8	80.8	79.8	79.5	80.3

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. ϕ is the parameter for AR(1) and AR(5) process. Bold numbers are significantly(at .05 level) less than .80.

Table 2.15: Simulated power (%) when working assumptions are true. Duration of the study is 6-week. The associated sample size is given in Table 2.11

Error Term	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
AR(1) $\phi = -0.6$	Pattern 1	22	78.5	80.3	78.5	82.3	79.8	80.3
		29	78.7	80.8	80.0	**77.1**	79.5	77.9
		36	77.7	80.3	80.2	78.2	**77.4**	83.6
AR(1) $\phi = -0.3$	Pattern 1	22	77.9	79.0	79.6	80.0	77.8	80.4
		29	77.9	79.1	80.0	79.0	78.0	78.4
		36	78.1	81.2	80.2	80.7	80.9	78.4
AR(1) $\phi = 0.3$	Pattern 1	22	80.2	78.5	80.8	80.5	79.6	82.6
		29	78.0	80.0	80.0	78.0	79.4	80.1
		36	77.6	82.5	80.6	77.0	78.9	82.0
AR(1) $\phi = 0.6$	Pattern 1	22	80.4	79.8	79.5	80.7	79.5	82.0
		29	78.9	81.5	79.3	79.5	81.3	79.5
		36	79.5	78.4	78.8	80.1	77.9	77.8
AR(5) $\phi = -0.6$	Pattern 1	22	79.9	79.4	80.0	78.7	79.2	79.4
		29	80.0	78.3	79.1	**76.8**	79.6	79.3
		36	80.5	80.0	79.2	80.1	78.0	80.4
AR(5) $\phi = -0.3$	Pattern 1	22	79.2	80.4	81.9	81.3	77.7	79.1
		29	80.0	82.3	80.5	80.5	82.2	79.2
		36	**75.9**	78.7	79.3	79.0	79.4	79.9
AR(5) $\phi = 0.3$	Pattern 1	22	79.4	80.8	79.8	79.5	**77.3**	81.2
		29	78.0	79.2	79.2	79.2	80.5	78.4
		36	78.3	79.1	78.1	80.7	80.5	79.5
AR(5) $\phi = 0.6$	Pattern 1	22	80.2	77.9	80.3	78.6	78.4	80.3
		29	**76.9**	79.3	80.2	79.1	80.6	80.5
		36	78.7	84.0	80.1	78.8	79.3	78.8

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. ϕ is the parameter for AR(1) and AR(5) process. Bold numbers are significantly(at .05 level) less than .80.

Table 2.16: Simulated type 1 error rate (%) when the duration of study is 4-week and 8-week. Error terms follow i.i.d. N(0,1). The associated sample size is given in Table 2.11.

Study Duration	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
4-week	Pattern 1	15	4.1	4.7	6.3	5.3	5.5	5.6
		22	5.2	4.4	4.7	3.1	4.7	4.4
		29	5.7	5.5	5.6	4.3	4.2	4.2
	Pattern 2	15	4.8	4.8	5.0	5.0	5.2	5.3
		22	5.1	5.2	4.7	3.7	4.2	3.7
		29	5.6	5.1	4.2	4.2	4.9	4.4
	Pattern 3	15	4.7	5.0	4.6	6.1	5.3	5.1
		22	4.9	4.0	**6.6**	4.2	3.8	4.1
		29	4.7	4.3	5.1	4.6	5.8	3.5
	Pattern 4	15	4.9	4.6	4.8	3.0	5.9	3.8
		22	3.5	5.1	4.5	5.2	3.8	6.0
		29	4.4	6.4	4.7	4.4	4.3	4.7
8-week	Pattern 1	29	4.1	4.6	4.0	5.3	5.0	5.9
		36	3.3	4.7	**6.5**	4.6	5.4	4.3
		43	3.2	5.1	5.2	5.0	3.4	5.0
	Pattern 2	29	3.9	5.0	4.5	4.2	3.7	4.1
		36	3.8	4.6	4.9	4.5	3.4	5.2
		43	3.9	5.4	5.0	3.4	3.8	5.0
	Pattern 3	29	4.6	4.2	3.7	5.2	4.1	4.0
		36	4.3	5.1	6.1	4.6	5.0	4.6
		43	4.6	6.0	4.1	5.0	4.9	4.0
	Pattern 4	29	4.5	5.2	2.9	3.6	5.3	4.4
		36	4.5	5.2	3.7	2.7	3.7	4.7
		43	4.2	**7.1**	4.9	4.4	4.5	4.8

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. Bold numbers are significantly(at .05 level) greater than .05 (for type I error)and less than 0.80 (for power).

Table 2.17: Simulated power (%) when the duration of study is 4-week and 8-week. Error terms follow i.i.d. N(0,1). The associated sample size is given in Table 2.11.

Study Duration	Availability Pattern	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7		
			Average Proximal Effect					
			0.10	0.08	0.06	0.10	0.08	0.06
4 week	Pattern 1	15	80.4	79.0	78.5	79.6	82.8	80.3
		22	78.8	78.7	80.7	78.7	79.2	80.0
		29	**76.2**	80.6	80.1	81.3	80.1	79.1
	Pattern 2	15	82.4	77.8	**77.2**	**75.9**	80.0	78.9
		22	**77.2**	80.3	81.5	**75.8**	80.7	82.0
		29	80.1	79.3	80.1	78.0	77.7	**76.9**
	Pattern 3	15	79.3	79.8	79.2	79.1	**76.5**	80.8
		22	80.0	80.0	79.0	79.0	80.2	81.8
		29	79.4	80.7	79.3	80.4	79.6	79.2
	Pattern 4	15	82.6	78.3	79.2	80.5	80.0	79.5
		22	80.4	80.7	79.3	79.1	78.5	79.2
		29	78.4	79.2	78.5	79.6	79.2	80.5
8 week	Pattern 1	29	79.7	**77.3**	**76.4**	79.1	82.2	79.6
		36	78.8	78.6	81.5	80.3	78.2	79.6
		43	80.4	77.8	78.7	79.1	80.3	80.1
	Pattern 2	29	79.3	81.1	79.8	78.7	79.7	80.2
		36	81.2	78.5	79.0	81.3	80.8	78.2
		43	80.3	81.5	77.5	**75.1**	78.8	78.1
	Pattern 3	29	80.1	79.0	**77.1**	78.2	80.4	78.8
		36	79.5	79.9	79.6	80.0	80.8	79.6
		43	80.5	79.5	79.6	79.4	79.4	80.2
	Pattern 4	29	82.1	79.7	80.7	79.7	79.0	78.4
		36	77.8	78.2	80.1	77.9	**76.9**	79.5
		43	79.6	78.5	78.1	79.4	80.6	79.5

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T}\mathbb{E}[I_t]$ is the average availability. Bold numbers are significantly(at .05 level) greater than .05 (for type I error)and less than 0.80 (for power).

Table 2.18: Simulated Type I error rate (%) and power (%) when the availability indicator, I_t depends on the recent past treatments with $\eta = -0.2$. The expected availability is constant in t and equal to 0.5. Duration of study is 42 days. The associated sample size is given in Table 2.11.

Error Term	Max	$\bar{\tau}$ = 0.5			$\bar{\tau}$= 0.7			$\bar{\tau}$ = 0.5			$\bar{\tau}$ = 0.7		
		Average Proximal Effect											
		0.10	0.08	0.06	0.10	0.08	0.06	0.10	0.08	0.06	0.10	0.08	0.06
AR(1) $\phi = -0.6$	22	4.8	5.4	4.5	3.4	5.8	3.7	81.5	78.0	79.4	81.7	77.9	80.7
	29	4.7	4.4	4.2	4.0	4.9	4.6	79.4	80.9	80.7	78.2	79.2	79.7
	36	4.3	5.3	4.4	4.2	3.9	5.5	79.5	81.5	79.8	80.2	79.2	80.7
AR(1) $\phi = -0.3$	22	4.7	3.8	4.4	3.5	4.4	4.6	78.7	81.2	80.3	80.9	77.9	78.5
	29	3.8	4.0	4.9	3.5	5.0	4.4	80.1	79.5	81.2	**77.3**	79.5	**77.1**
	36	2.7	5.7	4.0	3.3	4.7	5.2	**76.8**	80.4	79.9	78.8	79.5	79.4
AR(1) $\phi = 0.3$	22	4.8	4.1	4.4	5.0	5.4	3.6	83.0	79.8	79.4	81.3	78.9	79.2
	29	4.9	4.6	5.0	4.4	5.5	5.6	79.5	80.3	82.2	78.5	80.7	77.6
	36	4.9	4.9	4.2	3.3	4.5	4.8	80.0	78.9	79.5	81.7	79.4	79.6
AR(1) $\phi = 0.6$	22	4.5	5.1	4.7	4.3	4.6	4.0	80.3	78.9	81.1	81.2	81.5	77.9
	29	3.4	4.5	5.1	4.4	4.3	4.6	79.3	**76.2**	79.4	81.3	80.6	79.4
	36	4.8	4.3	4.2	4.1	4.5	4.5	77.5	80.5	80.9	76.7	80.0	79.7
AR(5) $\phi = -0.6$	22	4.8	4.6	4.3	3.7	4.7	3.5	81.9	81.4	81.6	79.8	78.3	78.9
	29	6.5	4.1	4.5	3.3	4.5	4.8	77.5	79.9	79.8	79.9	79.3	79.3
	36	3.5	5.7	4.4	4.6	4.7	5.7	77.8	80.8	78.6	77.9	79.2	81.7
AR(5) $\phi = -0.3$	22	4.3	4.9	4.0	4.3	5.6	5.0	77.7	81.8	80.0	80.1	80.3	81.1
	29	3.9	4.0	5.0	3.2	5.7	5.1	80.0	80.9	80.3	80.6	80.3	77.8
	36	4.0	3.6	4.7	4.8	4.8	3.2	79.0	80.4	80.8	80.1	79.0	**76.5**
AR(5) $\phi = 0.3$	22	3.5	4.9	5.0	4.1	3.8	4.1	**77.4**	82.9	78.5	80.6	81.4	80.2
	29	4.6	6.1	4.7	4.7	4.1	4.1	78.7	82.0	78.0	81.4	**76.5**	81.3
	36	5.1	4.4	4.0	3.2	3.9	4.7	79.7	81.8	78.6	79.1	**77.4**	79.0
AR(5) $\phi = 0.6$	22	5.0	4.6	4.3	4.0	4.0	5.5	80.5	79.4	82.5	79.2	81.1	81.0
	29	5.6	4.3	6.9	5.6	3.4	3.1	78.3	80.0	80.5	80.8	80.4	78.4
	36	4.8	4.8	4.8	3.5	3.7	5.5	78.2	80.5	80.3	77.6	80.5	79.1

"Max"is the day in which the maximal proximal effect is attained. $\bar{\tau} = (1/T)\sum_{t=1}^{T} \mathbb{E}[I_t]$ is the average availability. ϕ is the parameter for AR(1) and AR(5) process. Bold numbers are significantly(at .05 level) greater than .05 and less than 0.80.

Table 2.19: Simulated type I error rate (%) and power (%) when working assumption (a) is violated. Scenario 1. The average availability is 0.5. The day of maximal proximal effect is 29.

θ	$\bar{d}$	Availability Pattern							
		Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 1	Pattern 2	Pattern 3	Pattern 4
$0.5\bar{d}$	0.10	5.5	4.6	4.2	5.1	79.7	79.4	80.5	80.1
	0.08	5.1	4.4	5.4	4.6	80.4	78.9	80.4	78.7
	0.06	4.1	5.5	4.6	4.3	77.5	82.7	81.0	81.0
$\bar{d}$	0.10	4.8	4.3	3.7	4.1	79.3	78.3	77.8	79.4
	0.08	5.4	4.9	4.6	5.5	78.8	79.3	78.0	80.6
	0.06	4.4	3.5	5.1	4.6	78.4	79.3	79.0	80.4
$1.5\bar{d}$	0.10	4.4	4.1	4.4	4.8	78.3	80.5	78.4	79.9
	0.08	5.0	4.3	4.3	3.9	80.5	79.7	78.7	81.9
	0.06	4.0	5.1	5.5	5.6	**77.2**	80.8	81.6	80.3
$2\bar{d}$	0.10	4.1	3.8	5.0	5.5	77.7	78.8	79.0	78.4
	0.08	4.0	5.0	3.7	5.7	79.3	81.5	79.1	79.4
	0.06	4.9	4.3	5.2	5.3	80.8	79.0	77.5	80.9

$\bar{d} = (1/T)\sum_{t=1}^{T} Z_t^\top d$ is the average proximal effect. θ is the coefficient of W_t in $\mathbb{E}[Y_{t+1}|I_t = 1]$. Bold numbers are significantly (at .05 level) greater than .05 (for type I error) and lower than 0.80 (for power).

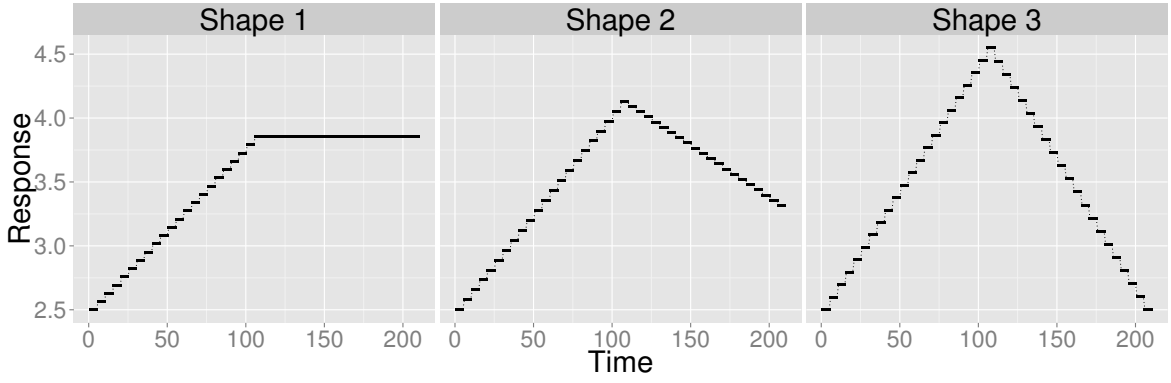

Figure 2.4: Conditional expectation of proximal response, $\mathbb{E}[Y_{t+1}|I_t = 1]$. The horizontal axis is the decision time point. The vertical axis is $\mathbb{E}[Y_{t+1}|I_t = 1]$.

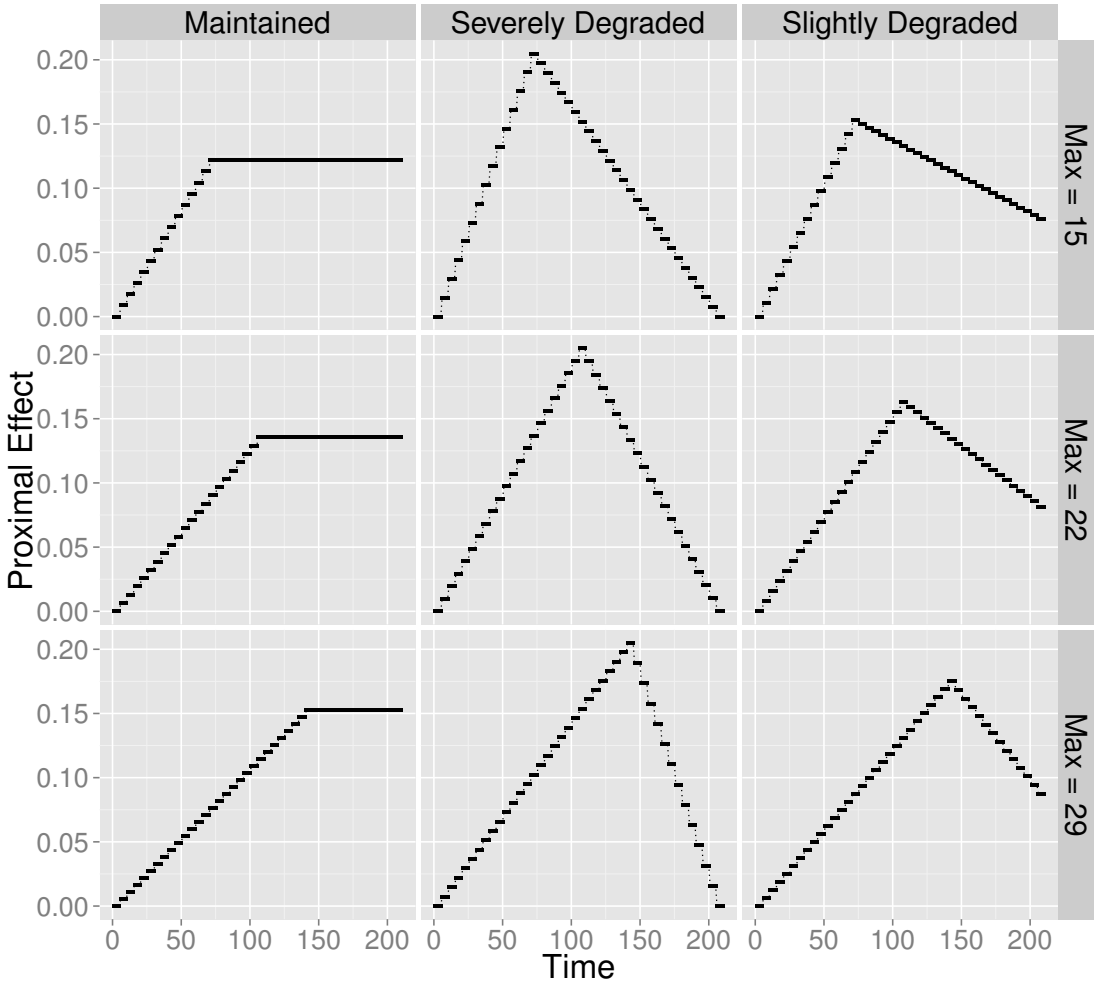

Figure 2.5: Proximal Main Effects of Treatment, $\{d(t)\}_{t=1}^{T}$: representing maintained, slightly degraded and severely degraded time-varying treatment effects. The horizontal axis is the decision time point. The vertical axis is the standardized treatment effect. The "Max" in the title refers to the day of maximal effect. The average standardized proximal effect is 0.1 in all plots.

Table 2.20: Degradation in power when average proximal main effect is underestimated. The day of maximal treatment effect is attained at day 29 and the average availability is 0.5 in all cases. The associated sample sizes for each value of average treatment effect are provided in first column.

$\bar{d}$ in Sample Size Formula	True $\bar{d}$	Availability Pattern			
		Pattern 1	Pattern 2	Pattern 3	Pattern 4
0.10 (N = 42)	0.098	76.2	78.9	77.6	78.6
	0.096	75.1	74.6	78.8	74.0
	0.094	73.7	70.7	75.4	73.4
	0.092	71.5	71.6	73.2	71.6
	0.090	68.9	68.4	69.6	67.3
	0.088	65.4	65.6	66.1	65.7
	0.086	66.4	67.9	65.2	66.7
	0.084	62.3	63.4	63.0	59.6
	0.082	60.0	60.2	60.5	58.2
	0.080	58.9	59.8	57.8	61.4
0.08(N = 64)	0.078	78.2	80.2	76.8	75.8
	0.076	77.3	76.7	76.2	75.4
	0.074	73.1	72.2	71.2	71.4
	0.072	70.7	71.0	69.4	68.2
	0.070	68.2	66.0	65.2	66.1
	0.068	65.5	64.3	64.6	65.7
	0.066	62.8	62.3	61.8	59.4
	0.064	61.9	58.5	59.5	62.1
	0.062	53.9	52.6	57.0	56.9
	0.060	54.6	51.1	54.8	53.4
0.06(N = 109)	0.058	75.6	76.9	74.0	78.1
	0.056	73.9	73.1	73.1	72.7
	0.054	68.6	71.1	69.3	68.5
	0.052	65.4	69.4	63.6	66.8
	0.050	61.0	62.8	64.1	63.2
	0.048	57.4	58.6	56.4	56.1
	0.046	53.6	53.4	52.9	54.8
	0.044	52.0	48.9	50.1	53.0
	0.042	45.7	43.9	44.9	46.4
	0.040	40.4	42.2	42.3	42.7

Table 2.21: Degradation in power when average availability is underestimated. The day of maximal treatment effect is attained at day 29 and the average proximal main effect is 0.1 in all cases. The associated sample sizes are given in first column.

$(1/T)\sum_{t=1}^{T}\tau_t$ in Sample Size Formula	True $(1/T)\sum_{t=1}^{T}\tau_t$	Availability Pattern			
		Pattern 1	Pattern 2	Pattern 3	Pattern 4
0.5 (N = 42)	0.048	76.4	81.7	76.0	78.2
	0.046	73.9	75.5	73.6	75.8
	0.044	70.6	72.1	71.0	71.7
	0.042	70.8	70.6	74.2	70.3
	0.040	70.3	69.2	65.7	68.6
	0.038	66.0	66.8	67.8	67.0
	0.036	64.0	62.5	62.4	62.9
	0.034	60.8	61.3	59.4	63.9
	0.032	56.4	59.2	54.7	59.8
	0.030	51.4	53.1	51.9	54.5
0.7 (N = 32)	0.068	79.5	76.1	79.1	75.0
	0.066	77.3	75.7	74.0	76.4
	0.064	74.5	74.7	73.5	77.1
	0.062	73.2	73.0	75.1	72.5
	0.060	69.8	70.5	73.5	72.5
	0.058	71.0	69.6	71.3	67.3
	0.056	68.8	70.3	66.6	64.0
	0.054	68.1	65.8	65.3	68.6
	0.052	62.4	64.9	65.6	62.9
	0.050	60.6	63.3	62.8	61.4

CHAPTER 3

Stratified Micro-Randomized Trails

Recent advances in mobile technologies have generated increased scientific interest in the use and development of Just-in-Time Adaptive Interventions (JITAIs) in mobile health. Wearable devices and/or smartphones can be used to unobtrusively collect data from users in real time, such as busyness, location, weather, step count and heart rate [34, 35, 36]. The JITAI involves treatments that are delivered via notifications on a smartphone or a wearable and which are designed to help users make healthy decisions to effectively manage their health and health behaviors. To be most effective in influencing health, the combination of both the right treatment and the right delivery time is likely critical [36]. Scientists are increasingly interested in designing mobile interventions in which treatments can be delivered to the user at the risk times, such as when the individual is stressed [37], anxious, or disengaging, and furthermore, in understanding whether it is useful to trigger delivery of treatments at these times.

In this chapter, we introduce *stratified micro-randomized trial* (sMRT) design, as a generalization or micro-randomized trial design in Chapter 2.2. This is motivated by our collaboration on the design of *Sense2Stop*, a mobile health smoking cessation study currently underway. In this study, participants are trained in stress reduction exercises prior to their smoking quit date. Apps that can be used to guide the participant through the exercises are installed on a study-provided phone. These apps can be accessed at any time by a participant. However, a common problem is that at the very times at which practicing these exercises might be most useful, participants do not do so. The scientific team is most interested in understanding whether reminders to practice stress-reduction exercises will be useful in reducing/preventing further stress if the reminders are delivered at stressful times. Thus, some reminders are to occur at times when the participant is classified as stressed (stress times) and the remaining at times the participant is not classified as stressed (not-stress times). A primary goal of this study is to assess whether the reminders result in a reduction/prevention of stress over the subsequent hour and how the effect differs between stress and non-stress times.

In the following section, we introduce Sense2Stop. Next, we introduce the stratified micro-randomized trial (sMRT) and discuss how sMRT generalizes the micro-randomized trial. We then define the causal treatment effect and construct a test statistic for assessing the treatment effect. Subsequently, we develop a simulation-based method for determining the sample size.

3.1 Sense2Stop: Smoking Cessation Study

Sense2Stop is a mobile health intervention study beginning on each participant's smoking quit day. This study includes a 10-day post-smoking-quit phase in which participant may receive reminders to practice self-regulation exercises installed on the smart phone [38]. These exercises are designed to help users manage their stress as stress is a risk factor for relapse to smoking. Participants wear both an AutoSense chest band [39] as well as bands on each wrist for 10 hours per day. Sensors in the chestband and wristband measure various physiological responses and body movements to robustly assess physiological stress. A online pattern-mining algorithm uses the resulting sensor data to construct binary time-varying stress classification at each minute of sensor wearing throughout the entire day [34].

3.2 Stratified Micro-Randomized Trial

Recall that an individual's longitudinal data, recorded via mobile devices that sense and provide treatments, can be written as

$$\{O_0, O_1, A_1, O_2, A_2, \ldots, O_t, A_t, \ldots, O_T, A_T, O_{T+1}\}$$

where t indexes decision times, O_0 is a vector of baseline information and $O_t (t \geq 1)$ is information collected between time $t-1$ and t. The treatment at time t is denoted by A_t; throughout this chapter we consider binary options for the treatments. In Sense2Stop, the decision time t is every minute during a 10 hour day over a period of 10 days and $A_t = 1$ if at decision t, the participant is prompted to practice stress-reduction exercises and $A_t = 0$ otherwise.

Similar to the micro-randomized trial (MRT) introduced in Chapter 2, the treatment in the *stratified micro-randomized trial* (sMRT) is randomized at each decision time. sMRT is a generalization of MRT to accommodate stratification. In particular, the decision times are divided into strata and the randomization occurs separately by strata. The primary rationale for the stratification is to ensure a sufficient number of decision times at which treatment is provided and is not provided

within one or within each of the strata. In particular, some strata occur more rarely and thus to ensure sufficient treatment exposure and non-exposure within each strata, we stratify randomization. For example, in Sense2Stop there are two strata: minutes at which a participant is classified as stressed and minutes at which the participant is not classified as stressed. Participants are expected to experience much fewer minutes of stress than non-stress minutes per day. The stratification variable at time t is denoted by $X_t \in \mathcal{X}$, included in the observation O_t. In Sense2Stop, $\mathcal{X} = \{0, 1\}$: $X_t = 1$ indicates that at the decision time t the participant is classified as stressed and $X_t = 0$ otherwise.

At some decision times, feasibility and ethics considerations imply that the participant is unavailable for treatment. For example, if sensors indicate that the user might be driving a car [34], then the message should not be sent; that is, the user is unavailable for treatment. The observation, O_t, includes an availability indicator I_t to capture this information; that is, $I_t = 1$ if the individual is "available" for treatment and $I_t = 0$ otherwise. In Sense2Stop, if a participant receives a treatment reminder, then for the next 60 minutes the participant is considered unavailable for further treatment. Furthermore in many settings the risk variable is defined using an interval of time. In Sense2Stop, the classification algorithm produces a smoothed probability of physiological stress across the minutes with an episodic pattern in which the probability increases, then decreases then increases and so on across the minutes. An episode is defined by the beginning of a positive trend interval and peaks at the end of a positive-trend interval followed by the start of a negative-trend interval. To ensure the required sensitivity and specificity the algorithm only attempt to make a classification right after the peak of an episode. Thus a participant can only be available for treatment in the minute after the peak of an episode.

Let $H_t = \{O_0, O_1, A_1, \ldots, A_{t-1}, O_t\}$ be the history of past treatments up time $t-1$ and observation up to time t. In sMRT, the treatment A_t is randomized with probability $\pi_t = \Pr(A_1 = 1|H_t)$. In the case where the participant is currently unavailable $I_t = 0$, the probability is set to 0, e.g., $\pi_t = 0$. Note that unlike in MRT, here the randomization probability $\pi_t = \Pr(A_t = 1|H_t)$ is allowed to depend on the stratification variable X_t to ensure we have enough treatments provided at each of strata. Furthermore, consideration of treatment burden often implies a constraint on the total number of treatments provided over certain time interval (e.g., in a day). Such "budget" constraint might result in a randomization probability that depends on the entire observed history, e.g., the number of provided treatments and the number of stress minutes in the current day. For example in Sense2Stop, the scientific team aims to provide an average of 1.5 reminders per day at stress times and not-stress times. In Chapter 4, we will discuss how to design the randomization probability to achieve the budget constraint by utilizing a forecasting function. Throughout we

assume the randomization probability $\pi_t = \Pr(A_t = 1|H_t)$ is known of history H_t. For ease of notation, denote the decision rule by $\pi_t(a|h_t) = \Pr(A_t = a|H_t = h_t)$

The mobile interventions are often designed to have proximal, near-term effects measured by certain proximal response variable. Typically the proximal response at time t is defined as a known function of $\{O_t, A_t, O_{t+1}\}$ and does not depend on the treatment at the next decision time, e.g., A_{t+1}. However, in the case where the decision times are dense in time (e.g., every minute), the proximal response might need to be defined over a subsequent time window. To distinguish between these two cases, we denote the proximal response at time t following the treatment A_t by $Y_{t,\Delta}$, a known function of $\{O_t, A_t, O_{t+1}, A_{t+1}, \ldots, O_{t+\Delta-1}, A_{t+\Delta-1}, O_{t+\Delta}\}$ over the subsequent time window of size Δ. Note that when $\Delta > 1$, the proximal response could be impacted future treatments. In Sense2Stop, the proximal response is the fraction of minutes classified as stress in the following one hour window, e.g., $Y_{t,\Delta} = \Delta^{-1} \sum_{s=1}^{\Delta} \mathbf{1}_{X_{t+s}=1}$ with $\Delta = 60$.

3.3 Proximal Main Effect of Treatment

In Section 2.3, Chapter 2, we defined the proximal effect of treatment when the proximal response is only a function of $\{O_t, A_t, O_{t+1}\}$, e.g. $\Delta = 1$. As discussed, when the size of time window $\Delta > 1$, the proximal outcome for time t can be impacted by future treatments, e.g., $\{A_{t+1}, \ldots, A_{t+\Delta-1}\}$, and thus the definition of proximal treatment effect involves careful examination of the distribution of future treatments. Below we define the proximal main effect of treatment using potential outcome framework [20, 21, 22].

Recall that the treatment can only be provided when the participant is unavailable and thus we need to index the potential outcome by decision rule. Recall in Section 2.3, we define $d(a, i)$ for $a \in \{0, 1\}$, $i \in \{0, 1\}$ by $d(a, 0) =$"unavailable-do nothing"and $d(a, 1) = a$. Then for each $a_1 \in \{0, 1\}$, define $D_1(a_1) = d(a_1, I_1)$. Then we denote the potential availability indicators at decision time 2 by $I_2^{D_1(a_1)} = I_2^{d(a_1,I_1)}$. Next for each $\bar{a}_2 = (a_1, a_2)$ with $a_1, a_2 \in \{0, 1\}$, define $D_2(\bar{a}_2) = d(a_2, I_2^{D_1(a_1)})$ and $\bar{D}_2(\bar{a}_2) = (D_1(a_1), D_2(\bar{a}_2))$. The potential availability indicator at decision time 3 is $I_3^{\bar{D}_2(\bar{a}_2)}$. Similarly, for each $\bar{a}_t = (a_1, \ldots, a_t)$, define $D_t(\bar{a}_t) = d(a_t, I_t^{\bar{D}_{t-1}(\bar{a}_{t-1})})$ and the sequence of decision rule up to time t, $\bar{D}_t(\bar{a}_t) = (D_1(a_1), \ldots, D_t(\bar{a}_t))$.

Now, the potential stratification variable at time t is then given by $\{X_t^{\bar{D}_{t-1}(\bar{a}_{t-1})}\}$. Recall that the proximal response $Y_{t,\Delta}$ is a function of $\{O_t, A_t, O_{t+1}, A_{t+1}, \ldots, O_{t+\Delta-1}, A_{t+\Delta-1}, O_{t+\Delta}\}$ and thus the potential proximal response is $Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{a}_{t+\Delta-1})}$. In this chapter, we consider the proximal

effect of treatment defined by

$$\begin{aligned}\beta(t,x) = \; & \mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1},1,\,0_{\Delta-1})} \mid I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, X_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = x\big] \\ & - \mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1},0,\,0_{\Delta-1})} \mid I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, X_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = x\big]\end{aligned}$$

where the expectation is taken with respect to the distribution of the potential outcomes and randomization in $\bar{A}_{t-1}$ according to the randomization probabilities $\{\pi_s = \Pr(A_s = 1|H_s), s \leq t-1\}$ and the future treatments in the potential proximal response are both set to $0_{\Delta-1} = (0, \ldots, 0) \in \mathbb{R}^{\Delta-1}$. Note that, similar to the proximal treatment effect defined in Section 2.3, the above proximal effect is conditional in that the effect of treatment at time t is defined for only individuals available for treatment at time t, that is, $I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1$ and the stratification variable at the level x, $X_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = x$. This proximal effect is a main effect in that the effect is marginal over any effects of previous treatment $\bar{A}_{t-1}$.

Here we set the future treatments $(a_{t+1}, \ldots, a_{t+\Delta-1}) = 0_{\Delta-1}$ in assessing the proximal treatment effect at time t. In Sense2Stop, the corresponding proximal effect can be interpreted as *the effect on the fraction of time stressed in the next hour of (a) providing a reminder at time t to practice stress-reduction exercises and no reminders within the next hour versus (b) no reminder at time t as well as within the next hour.* Other choice of future treatments is possible depending on the application and can be easily generalized to the stochastic case, i.e., the future treatments are selected according to some pre-specified decision rules. In particular, one can assume the future treatments are selected by the same randomization probabilities that select the treatments in the study as considered in [40]; similar to how we handle the distribution of previous treatments. However, this could be probabilistic if the randomization probability heavily depends on the current treatment. For example in Sense2Stop, the participant is unavailable in the next one hour after a reminder is sent. That is, the randomization probabilities for future treatments are all 0 when $A_t = 1$. This is very different comparing to the case when $A_t = 0$ so the proximal effect defined in this way is problematic. For the rest of chapter, we only consider the case where the future treatments are set to 0, but the generalization is straightforward.

Next, we express the proximal treatment effect $\beta(t, x)$ in terms of the observable data distribution. The expression is more complicated than in Section 2.4 in standard MRT due to the fact that (1) the randomization probability might depend on the entire history and (2) the proximal effect is defined by setting future $\Delta - 1$ treatments to 0, while in the observed data the treatments are

randomized. Assuming the consistency assumption [21, 22], the micro-randomization implies that

$$\beta(t,x) = \mathbb{E}\big[\mathbb{E}[W_t Y_{t,\Delta} \mid H_t, A_t = 1] - \mathbb{E}[W_t Y_{t,\Delta} \mid H_t, A_t = 0] \mid I_t = 1, X_t = x\big]$$

where the weight $W_t = \prod_{j=1}^{\Delta-1} \frac{\mathbb{1}_{\{A_{t+j}=0\}}}{1-\pi_{t+j}}$ and $\pi_j = \Pr(A_j = 1|H_j)$ is the randomization probability. To see this, note that for $a \in \{0,1\}$,

$$\begin{aligned}
&\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1}, a,\, 0_{\Delta-1})} \mid I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, X_t^{\bar{D}_{t-1}(\bar{A}_{t-1})}\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1}, a,\, 0_{\Delta-1})} \mid H_t(\bar{A}_{t-1})\big] \mid I_t^{\bar{D}_{t-1}(\bar{A}_{t-1})} = 1, X_t^{\bar{D}_{t-1}(\bar{A}_{t-1})}\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1}, a,\, 0_{\Delta-1})} \mid H_t\big] \mid I_t = 1, X_t\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t-1}, a,\, 0_{\Delta-1})} \mid H_t, A_t = a\big] \mid I_t = 1, X_t\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_t,\, 0_{\Delta-1})} \mid H_t, A_t = a\big] \mid I_t = 1, X_t\big]
\end{aligned}$$

where the second equality follows from the consistency assumption and the third equality follows from the randomization of A_t. On the other hand, the inside conditional expectation can be re-written using importance weights:

$$\begin{aligned}
&\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_t,\, 0_{\Delta-1})} \mid H_t, A_t\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_t,\, 0_{\Delta-1})} | H_{t+1}\big] \mid H_t, A_t\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\frac{\mathbb{1}_{\{A_{t+1}=0\}}}{1-\pi_{t+1}} Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_t,\, 0_{\Delta-1})} | H_{t+1}\big] \mid H_t, A_t\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\frac{\mathbb{1}_{\{A_{t+1}=0\}}}{1-\pi_{t+1}} Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t+1},\, 0_{\Delta-2})} | H_{t+1}\big] \mid H_t, A_t\big] = \cdots \\
&= \mathbb{E}\big[\mathbb{E}\big[\prod_{j=t+1}^{t+\Delta-1} \frac{\mathbb{1}_{\{A_j=0\}}}{1-\pi_j} Y_{t,\Delta}^{\bar{D}_{t+\Delta-1}(\bar{A}_{t+\Delta-1})} | H_{t+\Delta-1}\big] \mid H_t, A_t\big] = \mathbb{E}\big[W_t Y_{t,\Delta} \mid H_t, A_t\big]
\end{aligned}$$

where in the last equality we use the consistency assumption.

3.4 Test Statistic and Sample Size Calculation

Our sample size formula is based on a test statistic for use in testing

$$\mathbf{H}_0 : \beta(t,x) = 0,\ t = 1, \ldots T, x \in \mathcal{X}$$

against a scientifically plausible alternative. Similar to the test statistics developed in MRTs in Section 2.4, here we also develop the test statistics for use in testing $\mathbf{H}_0$ against alternatives of the linear form $f_t(x)^\top \beta$, where $f_t(x) \in \mathbb{R}^p$, is a feature vector of t and stratification variable x.

We base the test statistic on the estimator of β in a weighted least squares fit of a working model. In particular, define the $q \times 1$ vector $B_t = g_t(H_t)$, where g_t is a function of the current history H_t (including the stratification variable X_t) that are predictive of the proximal response. Define the $q \times 1$ vector $Z_t = f_t(X_t)$. Let $\tilde{\pi}_t(a|x)$ be a pseudo decision rule depending only on the stratification variable, that is, $\tilde{\pi}_t(a|x) \in [0,1]$ and $\sum_a \tilde{\pi}_t(a|x) = 1$ for each $x \in \mathcal{X}$, and let $\tilde{\pi}_t = \tilde{\pi}_t(1|X_t)$. The estimator $(\hat{\alpha}, \hat{\beta})$ minimizes the weighted least squared objective function with centered treatments:

$$\mathbb{P}_N \left\{ \sum_{t=1}^{T} I_t W_t \frac{\tilde{\pi}(A_t|X_t)}{\pi_t(A_t|H_t)} \left(Y_{t+\Delta} - B_t^\top \alpha - (A_t - \tilde{\pi}_t) Z_t^\top \beta \right)^2 \right\} \tag{3.1}$$

Recall the first weight is given by $W_t = \prod_{j=1}^{\Delta-1} \frac{\mathbb{1}_{\{A_{t+j}=0\}}}{1-\pi_{t+j}}$. The use of W_t is to adjust the discrepancy between the selection of future treatments in the study and in the definition of proximal treatment effect. The use of second weight $\frac{\tilde{\pi}(A_t|X_t)}{\pi_t(A_t|H_t)}$, motivated by [40], allows the consistent estimate of treatment when the proximal treatment effect model is correctly specified, e.g., $\beta(t,x) = f_t(x)^\top \beta$ for some $\beta \in \mathbb{R}^p$ even when the working model $B_t^\top \alpha$ for $\mathbb{E}[W_t Y_{t+\Delta}|H_t]$ is mis-specified. Similar to the least squared estimator presented in Section 2.4, the action-centering, e.g., $A_t - \tilde{\pi}_t$, guarantees that even when the treatment effect model is wrongly specified, the estimator $\hat{\beta}$ converges to

$$\tilde{\beta} = \left(\mathbb{E}\big[\sum_{t=1}^{T} I_t \tilde{\pi}_t (1-\tilde{\pi}_t) Z_t Z_t^\top \big] \right)^{-1} \mathbb{E}\big[\sum_{t=1}^{T} I_t \tilde{\pi}_t (1-\tilde{\pi}_t) \beta(t, X_t) Z_t \big] \tag{3.2}$$

Note that $\tilde{\beta}(t,x) = f_t(x)^\top \tilde{\beta}$ is the weighted L_2 projection of the proximal treatment effect onto the linear space spanned by $f_t(x)$. Similarly, one can show that $\hat{\alpha}$ converges to $\tilde{\alpha}$:

$$\tilde{\alpha} = \left(\mathbb{E}\big[\sum_{t=1}^{T} I_t B_t B_t^\top \big] \right)^{-1} \mathbb{E}\big[\sum_{t=1}^{T} I_t \big(\sum_a \tilde{\pi}(a|X_t) \mathbb{E}[W_t Y_{t+\Delta}|H_t, A_t = a] \big) B_t \big]$$

Furthermore, the weighted least squares estimators $\hat{\beta}$, under moment and invertibility conditions, satisfies that $\sqrt{N}(\hat{\beta} - \tilde{\beta})$ is asymptotically normal with mean 0 and variance $\Sigma_\beta = Q^{-1} W Q^{-1}$

where $Q = \mathbb{E}\big[\sum_{t=1}^{T} I_t \tilde{\pi}_t(1-\tilde{\pi}_t) Z_t Z_t^\top\big]$,

$$W = \mathbb{E}\left[\Big(\sum_{t=1}^{T} \tilde{\epsilon}_t I_t W_t \frac{\tilde{\pi}(A_t|X_t)}{\pi_t(A_t|H_t)}(A_t - \tilde{\pi}_t) Z_t\Big) \times \Big(\sum_{t=1}^{T} \tilde{\epsilon}_t I_t W_t \frac{\tilde{\pi}(A_t|X_t)}{\pi_t(A_t|H_t)}(A_t - \rho_t) Z_t^\top\Big)\right]$$

and $\tilde{\epsilon}_t = Y_{t+\Delta} - B_t^\top \tilde{\alpha} - (A_t - \tilde{\pi}_t) Z_t^\top \tilde{\beta}$. Using the small sample adjustment as discussed in Section 2.4, the rejection region for the test of $\mathbf{H}_0$ is then

$$\left\{ N \hat{\beta}^\top \hat{\Sigma}_\beta^{-1} \hat{\beta} > \frac{p(N-q-1)}{N-q-p} F^{-1}_{p,N-q-p}(1-\alpha_0) \right\} \tag{3.3}$$

where $\hat{\Sigma}_\beta^{-1}$ is the sandwich estimator and α_0 is the significance level. The proof of this result is similar to Lemma 1 in Chapter 2 and thus omitted here.

To determine the sample size for sMRT, we aim to calculate the smallest sample size needed to detect an alternate $\beta(t,x)$ with a given power $(1-\beta_0)$ at a given significance level (α_0). When N is large and $\mathbf{H}_1$ holds, $N\hat{\beta}'\hat{\Sigma}_\beta^{-1}\hat{\beta}$ is approximately distributed as a noncentral chi-squared $\chi_p^2(c_N)$ [26], where the non-centrality parameter $c_N = N\tilde{\beta}^\top \Sigma_\beta^{-1} \tilde{\beta}$ and $\tilde{\beta} = Q^{-1} W Q^{-1}$ is defined in (3.2). With the access to c_N, the desired sample size can be found according to (2.9). Unfortunately, calculation of noncentrality parameter, c_N in sMRT is non-trivial due to the fact the randomization probability depends on the stratification variable or even the entire history. In order to find an analytic form of the noncentrality parameters, one need to impose certain strong assumption on the distribution of the stratification variable besides the working assumptions presented in Section 2.5. As such, we below propose a three-step simulation-based sample size calculator.

In the first step, information elicited from the scientist is used to formulate a generative model and calculate, via Monte-Carlo integration, $\gamma_c = \tilde{\beta}^\top \Sigma_\beta^{-1} \tilde{\beta}$ in the non-centrality parameter. The resulting value, $\hat{\gamma}_c$, is plugged in to equation 2.9 to solve for an *initial* sample size $\hat{N}_0$. In the second step, we use a binary search algorithm to search over a neighborhood of $\hat{N}_0$; in our simulations, we found the binary search quickly resulted in a solution. For each sample size N required by the binary search algorithm, K samples each of N simulated participants are run. Within each simulation, the rejection region for the test is given by equation (3.3) at the specified significance level. The average number of rejected null hypotheses across the K simulations is the estimated power for the sample size N. The sample size is the minimal N with estimated power above the pre-specified threshold $1-\beta_0$. In the last, third, step we conduct a variety of simulations to assess the robustness of the sample size calculator to any assumptions and to make adjustments to ensure robustness.

CHAPTER 4

Determining Treatment Timing Under Average Constraint

The Just-In-Time Adaptive Interventions (JITAIs) involves treatments which are designed to help users make healthy decisions in the moment. To be most effective in influencing health, the combination of both the right treatment and right delivery time is likely critical [36]. Many prediction/detection algorithms have been developed in order to provide potential delivery times. For example, physiological measurements collected from wearable sensors can be used to detect physiological stress [39]. Impending negative mood or detections of current negative mood [41] and detections of risky locations [42, 43] are additional examples of risk predictions/detections. Alternatively, the algorithms may detect times of potential receptivity or interruptibility [44].

Scientists are increasingly interested in designing mobile interventions in which treatments can be delivered to the user at these times. For example, in smoking cessation, we might aim to deliver a reminder to practice stress management skills when the user is detected to be stressed in order to limit potential relapse. Depending on the user and the day, there may be many such times. However, it is well understood that delivering too many treatments can cause undue user burden [45], possibly leading to app disengagement. Furthermore, repeatedly providing similar treatments may lead to habituation, where users begin to pay less attention to each subsequent treatment, decreasing its effectiveness [36, 46]. Thus, scientists often impose constraints on the number of times the mobile device should deliver treatments. Consider, for instance, the HeartSteps V1 physical activity study [47, 48, 49] in which activity suggestions are delivered to the user's phone. In this study, considerations of burden and habituation led to the constraint that, on average, three activity suggestion messages would be delivered per day. In the planned next version of HeartSteps – referred to as HeartSteps V2 throughout – one of the treatment components is an anti-sedentary message. Here, the scientific team aims to provide an average of 1.5 anti-sedentary messages per day at sedentary times.

Here, times at which treatment may be provided are referred to as risk times. Ideally, it is best to deliver the treatments uniformly across the risk times so as to randomly sample the full variety of contexts in which risk times occur. Uniform sampling benefits the study design in two ways: first, uncertainty in when the treatments are delivered can reduce user habituation [46]; second, the uniform sampling of risk times enhances the ability of data analyses to learn if and in which contexts there is a causal effect of the treatment.

In this chapter we develop a "**Seq**uential **R**isk **T**ime **S**ampling" (SeqRTS) algorithm that both satisfies the desired constraint on the total number of treatments in a given time interval and spreads these treatments uniformly across all risk times. The SeqRTS algorithm combines forecasts of the remaining number of risk times within future blocks of time with a sequential algorithm that, at each risk time, provides a probability for triggering delivery of treatment. The proposed algorithm can be used to design the randomization scheme for interventions in *stratified micro-randomized trials* design in Chapter 3

This work is motivated by our collaboration on two mHealth studies – a smoking cessation trial currently in the field, Sense2Stop [35, 50], and in planning the next version of a physical activity trial, HeartSteps V2 [48]. In both cases, the approach developed here is currently being used or will be used to sample risk times to provide treatment. In both of the motivating studies, a primary goal is to learn if and in which contexts there is a causal effect of the treatment in altering health behaviors. In HeartSteps V2, for example, one goal is to determine if the anti-sedentary messages are effective at times the user is detected to be sedentary and how this effectiveness might be impacted by current context. In the smoking cessation study, Sense2Stop, one goal is to learn whether the reminders to practice stress management skills are effective at times the user is detected to be stressed.

4.1 Related work

A natural approach to developing a method to meet the constraints and deliver treatments uniformly across risk times is to build on methods from the ecological momentary assessment (EMA) [51, 52, 53, 54, 55, 56] literature. Recall an EMA is a self-report collected via a mobile device as the user goes about his/her life [51, 56]. However, due to the high user burden imposed by frequent requests for self-reports, scientists often set a budget for the number of EMA requests within a day. Indeed, a higher average EMA response rate is observed in nonclinical studies when users are prompted for self-report fewer times per day [57]. In addition, usually scientists aim to uniformly spread out the EMA data collection across the day so that the self report answers more accurately

reflect the user's mood/behaviors in different contexts throughout the day.

A classical approach for timing the EMA is to split each day into some number of blocks, say K, and assign each block with certain number of treatments to achieve the constraint [52, 53, 54]. In a recovery support services study [52], for example, the day was split into $K = 5$ blocks. Within each block, a time was uniformly sampled and an EMA was delivered via the mobile device to the user at the sampled time. This method achieves the budget constraint exactly of five EMA messages every day. Of course, the number of messages can be randomized. For instance, if we want to achieve an average of 3 messages per day and we keep the 5 blocks from the prior example, then we can send a message with probability $3/5 = 0.6$ in each block. If a block is selected for a message then the time at which the EMA message is sent is sampled uniformly within the block.

Rathbun et al. [55] consider alternative approaches to sending EMA. They sample times at which to send EMA according to a Poisson process with intensity $\lambda_t(\mathcal{H}_t) = \lim_{\delta \to 0} \delta^{-1} \mathbb{E}[N[t, t + \delta)|\mathcal{H}_t]$, where $N[a, b)$ is the number of EMA sent within the time interval $[a, b)$ and $\mathcal{H}_t$ denotes history of all the EMA times before time t. One option they consider for the intensity is $\lambda_t(\mathcal{H}_t) = \exp(\alpha + \beta(t - \rho N[0, t)))$ for some $\beta, \rho > 0$. This intensity self-corrects when the system has sent more EMA than was desired. That is, if $N[0, t)$, the number of EMA sent prior to time t, goes well-above the target t/ρ then the probability of sending an EMA is decreased.

However, these methods were not developed to deliver treatment. The method developed in this paper generalizes ideas from the above methods to the setting in which EMA messages are only to be sent at risk times and in addition, we do not know how many risk times will occur within any block of time. We will see that, when there is high variability in the number of risk times within the block, the SeqRTS algorithm outperforms simple extensions of the block sampling approach in achieving the desired average number of treatments and in spreading the treatments uniformly across risk times. We will also see that this performance depends on the forecast quality.

In the next section, we introduce notation for the longitudinal data collected from wearable devices. Next, in section 4.2, we introduce the SeqRTS algorithm. We discuss each tuning parameter, what it controls, and how to set their values. We evaluate the performance of the SeqRTS algorithm in two mHealth studies – the Minnesota smoking study and HeartSteps V1. Studying performance on the Minnesota study informs expected performance in Sense2Stop. We end with a discussion of limitations and suggestions regarding practical implementation in future studies.

4.2 Sequential Risk Time Sampling Algorithm

Data and Notation

We recall that the user's longitudinal data recorded via mobile devices can be written as

$$\{O_0, O_1, A_1, O_2, \ldots, O_t, A_t, O_{t+1}, \ldots\}$$

where t indexes regularly-spaced times (e.g., every minute, five-minutes, thirty-minutes, hour, etc.); O_0 contains the baseline information; the observation O_t ($t \geq 1$) is the vector of sensed and self-report observations collected between time $t-1$ and t; and A_t is the treatment at time t. Choice of time-scale is usually determined by the frequency with which the risk detections can be made. In the smoking cessation study, the temporal frequency is set to every minute; in the physical activity study, the frequency is every five minutes. For simplicity, we consider binary treatment, i.e., $A_t = 1$ if treatment is delivered, $A_t = 0$ otherwise. Denote by $H_t = \{O_0, O_1, A_1, O_2, \ldots, A_{t-1}, O_t\}$ the observation history up to time t as well as the treatment history at all times up to, but not including, time t.

The observations, O_t, include a variable that indicates risk, X_t. For example, in the physical activity study, described below, $X_t = 1$ if the user's wristband tracker records less than 150 steps in the past 40 minutes (i.e., user is sedentary) and $X_t = 0$ otherwise. The activity suggestion messages are designed to be delivered when the user is sedentary, e.g., when $X_t = 1$. A *risk time* is a time t at which the user is detected to be at risk. In the physical activity study, "at risk" implies $X_t = 1$; however, in other studies, multiple levels of risk may exist. In the smoking cessation study, for example, $X_t = 1$ (i.e., time is not classified as stressed) and $X_t = 2$ (time is classified as stressed) are two levels of risk. Write $X_t = 0$ to denote the user is not at risk, and $X_t = x \in \{1, \ldots, \mathcal{X}\} > 0$ to denote the user is at risk level x at time t (i.e., a risk time at level x). The risk variable at time t, X_t, is contained in H_t.

At some risk times, however, feasibility and ethics considerations imply that the individual is unavailable for treatment. For example, if sensors indicate that the user might be driving a car [34], then the message should not be sent; that is, the user is unavailable for treatment. The observation, O_t, includes an availability indicator I_t to capture this information; that is, $I_t = 1$ if the individual is "available" for treatment and $I_t = 0$ otherwise. An available time is a time t at which the user is available for treatment, i.e., when $I_t = 1$. The availability indicator at time t, I_t, is contained in H_t. An available risk time is a time t at which the user is available for treatment and at risk. Finally, times t such that $X_t = x > 0$ and $I_t = 1$ are referred to as available risk times at level x.

Algorithm

As mentioned above, the proposed sequential risk time sampling (SeqRTS) algorithm generalizes blocking as well as the use of a sequential algorithm from the EMA literature. For simplicity, assume that each day is split into K time blocks, i.e., $\mathcal{T} = \cup_{k=1}^{K} \mathcal{B}_k$ and denote the size of each block $|\mathcal{B}_k| = T$. The following method can be easily generalized to allow for different numbers of blocks per day or differently sized blocks depending on the time of day. Suppose there are multiple levels of risk indexed by $x \in \{1, \ldots, \mathcal{X}\}$. And suppose that each day, an average of N_x^* available risk times at level x are to be sampled for treatment delivery ($A_t = 1$). Mathematically, the average constraint can be written as

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}} A_t \mathbb{1}_{\{X_t = x, I_t = 1\}}\right] = N_x^* \tag{4.1}$$

for each value of x. Note the expectation is over the distribution of available risk times at level x within a given day. Furthermore, a secondary goal is to deliver treatment uniformly over available risk times at every level x such that the above constraint is satisfied. Operationally, the goal is to design a probability to assign treatment or, equivalently, a probability that is used to sample available risk times at level x.

Intuitively, at each available risk time in a block, SeqRTS calculates the number of remaining available risk times to be sampled for treatment in the block and divides this by the expected number of available risk times remaining in the block. The formula of randomization probability presented below is derived by iteratively taking the conditional expectation given the current history of (4.1); see Appendix 4.5 for the detailed derivation. Below we first introduce the algorithm and provide the intuition and then discuss the inputs required by the algorithm. Suppose time t is an available risk time in the k-th time block at level x for a user. Then, SeqRTS delivers treatment with probability

$$\pi_t = \pi_t(H_t) = \phi_\epsilon \left(\frac{N_{x,k} - C_{t,\lambda}(x)}{1 + g(x \mid H_t)} \right), \tag{4.2}$$

where

1. $N_{x,k} \geq 0$ (i.e., the block budget) is a tuning parameter. Roughly speaking, $N_{x,k}$ is the average number of treatments delivered in block k at level x. So $\sum_{k=1}^{K} N_{x,k} \approx N_x^*$.

2. $\lambda \in [0, 1]$ is a tuning parameter and $C_{t,\lambda}(x)$ denotes a soft version of number of treatments

that have been triggered so far in current block at the risk level x; that is,

$$C_{t,\lambda}(x) \;=\; \sum_{s\in\mathcal{B}_k, s\le t-1} \mathbb{1}_{\{I_s=1, X_s=x\}}\big(\lambda^{t-s}A_s + (1-\lambda^{t-s})\pi_s\big). \tag{4.3}$$

$C_{t,1}(x)$ (i.e., setting $\lambda = 1$) equals the exact number of treatments that have been triggered so far in the current block at risk level x. $C_{t,0}(x)$ (i.e., setting $\lambda = 0$) equals the sum of probabilities of triggering treatment at previous available risk times at level x in the current block. The former uses the observed history. The latter uses the "expected" history. The choice of $\lambda \in [0, 1]$ smoothly adjusts between these extremes. As will be seen below, the tuning parameter λ controls the variability in sampling the risk times.

3. $g(x|H_t)$ denotes a forecast of the number of available risk times at risk level x left in the current block, i.e., $\sum_{s\in\mathcal{B}_k, s\ge t+1} \mathbb{1}_{\{I_s=1, X_s=x\}}$, given the observed history up to time t, H_t.

4. $\phi_\epsilon(x) \;=\; x\mathbb{1}_{\{x\in[\epsilon_L,\epsilon_U]\}} + \epsilon_U\mathbb{1}_{\{x>\epsilon_U\}} + \epsilon_L\mathbb{1}_{\{x<\epsilon_L\}}$ denotes a truncation function with pre-specified upper and lower limits $\epsilon = \{\epsilon_L, \epsilon_U\}$. The truncation function ensures the output value stays within $[\epsilon_L, \epsilon_U]$ (i.e., bounded away from 0 and 1) to allow for causal inference with the collected data. This truncation function is intended as a last resort as generally $N_{x,k}$ and λ, when well tuned, will ensure that the fraction in (4.2) is bounded away from 0 and 1.

To recap, the numerator in (4.2) takes the block budget $N_{x,k}$ for level x and subtracts the amount that has been "used" by time t (i.e., $C_{t,\lambda}(x)$); this is, roughly speaking, the number of remaining times to be sampled for treatment in the block where the user will be available and at risk level x. This "remaining budget" is then divided evenly among the expected number of available risk times at level x remaining in the block (i.e., $1 + g(x \mid H_t)$). Algorithm 1 provides pseudocode for the SeqRTS algorithm at a particular risk time t.

In general, we aim to sample times for treatment with probabilities bounded away from 0 and 1; this enhances our ability to learn the casual effect of the treatment and how this causal effect is impacted by the user's context. However, if $\epsilon_L = 0$, $\epsilon_U = 1$ and $\lambda = 1$, then it is possible, given the history, for a time at which the user is available and at risk level x to be sampled for treatment with zero probability. This is because the numerator becomes zero whenever the past number of treatments in the current block equals the target $N_{x,k}$. Additionally, for certain histories, the probability of sending treatment can be one. Consider a toy example with a perfect forecast, the block is a day and every day has five available risk times. Suppose the goal is to achieve an average of one treatment per day. Then if $\lambda = 1$, once a treatment within the block is provided, the probability of treatment at any future available risk time in the day is zero. Or if no treatment

ALGORITHM 1: SeqRTS algorithm applied at a risk time t

Input: Current block budgets: $\{N_{x,k}\}_{x=1,\ldots,\mathcal{X}}$. Tuning parameter: λ. Current history: H_t, which contains current risk and availability (X_t, I_t). Forecasting Method: $g(x|h)$. Bounds: $\epsilon = (\epsilon_L, \epsilon_U)$

Output: Current treatment A_t and treatment probability π_t

Set $x \leftarrow X_t$

if $I_t = 0$ **then**

  Set $\pi_t = 0$ and $A_t = 0$

end

else

  Compute $C_{t,\lambda}(x)$ via (4.3)

  Compute $g(x|H_t)$ for given forecasting method

  Set $\pi_t = \phi_\epsilon\left(\frac{N_{x,k} - C_{t,\lambda}(x)}{1+g(x|H_t)}\right)$

  Draw $A_t \sim \text{Bern}(\pi_t)$

end

return $\{\pi_t, A_t\}$

has been provided in the first four risk times, then the probability of sending treatment at the fifth risk time becomes 1. To avoid these settings we select $\lambda < 1$ and, in addition, we employ the truncation function, ϕ_ϵ, with $\epsilon_L > 0$ and $\epsilon_U < 1$. Furthermore, if $\lambda = 0$, the algorithm, because it does not take into account past treatments but only the probabilities of past treatments, may sample many more or much fewer risk times than desired. Consider the toy example once more, in which the block is an entire day. Then if $\lambda = 0$, the average number of treatments per day is equal to 1 as desired, but on 19% of the days, the user will receive more than 3 treatments. For studies where treatment may cause high undue user burden, this might be considered to be excessive. In Sense2Stop, for instance, an excessive number of treatments at times classified as "Stressed" may only exacerbate stress. Tuning of λ guards against the likelihood of over-treating at these times. See section 4.5 for a more complex example of this trade-off.

When the scientific team believes the variation in the number of treatments under $\lambda = 0$ is too high, one could choose a non-zero λ to reduce the variation. Additionally, the use of discounted weights λ^{t-s}, instead of a fixed, time-invariant weight, is to help spread out the treatments, as the probability of sending treatments would decrease if a treatment was delivered in the recent past and such impact would be weaken as time goes on (discounted by the length of separation $t - s$). This is similar to the use of the self-correcting process in the point-process sampling method for EMA discussed in section 4.1.

A key difference between the SeqRTS algorithm and the block sampling method discussed in section 4.1 is the use of forecast, $g(x|H_t)$. Essentially, we replace the crude estimate of the average number risk times per block by time-varying forecasts of the remaining risk times at risk level x

within the block. These forecasts allow us to use user-specific time varying covariates and baseline characteristics to account for the potentially high variability in the number of risk times and in how the risk times are spread out within a block and thus better achieve the average constraint and uniformly spread out the treatments across the risk times.

Selecting Tuning Parameters SeqRTS requires selection of blocks, the construction of the forecasts ($g(x|H_t)$) and the tuning parameters λ and $\boldsymbol{N} = \{N_{x,k}, x = 1, \ldots, \mathcal{X}, k = 1, \ldots, K\}$. We assume that the blocks and bounds ϵ have been selected. Because there are a variety of high quality prediction/forecasting methods available, here we focus on tuning of λ and $\boldsymbol{N}$. See below for comments on the selection of the number of blocks as well as the prediction method. The TUNE algorthim is given in Algorithm 2.

Training data is used to tune λ and $\boldsymbol{N}$. This training data must include all of the features needed by the forecasting method as well as the risk X_t and availability I_t variables. Here, our training data is from similar studies to Sense2Stop for which the same sensing suites are deployed – the Minnesota smoking study.

As previously discussed, the value of $N_{x,k}$ controls the total number of treatments in the k-th block. When there is more than one time block in a day (i.e. $K > 1$), in order to choose an appropriate value of $N_{x,k}$ we first construct a target average number of treatments for each block, denoted by $N^*_{x,k}$, by splitting the overall daily constraint N^*_x into K time blocks such that $N^*_x = \sum_{k=1}^{K} N^*_{x,k}$. Here we use $N^*_{x,k} = N^*_x/K$.

To tune the parameters $(\lambda, \boldsymbol{N})$, we use (4.2) to determine the probability π_t to generate an A_t at each available risk time. This is done for each block in each day in the training data, 1000 times. We then compute the average number of treatments in the k-th block at each level x (across all days and the 1000 runs of the algorithm) and denote the averages by $F_{x,k}(\lambda, \boldsymbol{N})$. For each λ in a grid, we search for the optimal tuning value of $N_{x,k}$, such that the computed average number of treatments is equal to the target constraint $N^*_{x,k}$; more precisely, we minimize the objective function $J(\boldsymbol{N}) = \sum_{k=1}^{K} \sum_{x=1}^{\mathcal{X}} (N^*_{x,k} - F_{x,k}(\lambda, \boldsymbol{N}))^2$. The remaining problem is how to tune λ. Recall that we aim to select a value of $\lambda < 1$ so as to ensure the sampling probabilities lie in $(0, 1)$. However small values of λ can potentially result in too much variance in the number of treatments (e.g. sampled risk times) across days. Our approach, as part of the scientific team, is to decide what level of daily variation in treatments is tolerable and use this to tune λ. That is, we specify a probability, p, and a range $[l, u]$ so that the probability of total treatments within a given range $[l, u]$, i.e., $\Pr(l \leq \sum_{t \in \mathcal{T}} A_t \mathbb{1}_{\{I_t = 1, X_t = x\}} \leq u) \geq p$ for each level of risk $x \in \{1, \ldots, \mathcal{X}\}$. For each value of λ, the training data to estimate this probability under the optimal, tuned $\boldsymbol{N}$. Then

ALGORITHM 2: TUNE: finds tuning parameters for SeqRTS using training data

Input: Block budgets: $\{N^*_{x,k}\}$; Person-day risk and availability trajectories $\{(X_{i,t}, I_{i,t})\}_{i=1,t=1}^{n,T}$. Forecasting method $g(x|h)$. Bounds $\epsilon = \{\epsilon_L, \epsilon_U\}$. Grids G_λ and $G_{\boldsymbol{N}}$ for λ and $\boldsymbol{N}$. Treatment range $[l, u]$ and lower-bound probability p

Output: Tuning parameters: $(\lambda_{\text{opt}}, \boldsymbol{N}_{\text{opt}})$

for $(\lambda, \boldsymbol{N}) \in G_\lambda \times G_{\boldsymbol{N}}$ **do**
  for $i \leftarrow 1$ ***to*** n **do**
    for $j \leftarrow 1$ ***to*** 1000 **do**
      Initialize $H_0 = \emptyset$; $\pi_{i,0} = 0$; $A_{i,0} = 0$
      for $t \in \mathcal{T}$ **do**
        Set $H_{i,t} = \{H_{i,t-1}, \pi_{i,t-1}, A_{i,t-1}, X_{i,t}, I_{i,t}\}$
        Set $\{\pi_{i,t}, A_{i,t}\} \leftarrow$ SeqRTS with inputs $(\boldsymbol{N}, \lambda)$, $H_{i,t}$, $g(x|h)$, and ϵ
      end
      Store ith person-day, jth iteration history output as $H_{i,j}(\lambda, \boldsymbol{N})$
    end
  end
  for $x = 1, \ldots, \mathcal{X}$ *and* $k = 1, \ldots, K$ **do**
    Compute $F_{x,k}(\lambda, \boldsymbol{N})$ using $\{H_{i,j}(\lambda, \boldsymbol{N})\}_{i=1,j=1}^{n,1000}$
  end
end
for $\lambda \in G_\lambda$ **do**
  Set $\hat{\boldsymbol{N}}(\lambda) = \arg\min_{\boldsymbol{N} \in G_{\boldsymbol{N}}} \sum_{k=1}^{K} \sum_{x=1}^{\mathcal{X}} (N^*_{x,k} - F_{x,k}(\lambda, \boldsymbol{N}))^2$
  Compute $\hat{P}_{\lambda,x}$: empirical estimate of $\Pr(l \leq \sum_{t \in T} A_t \mathbb{1}_{\{I_t=1, X_t=x\}} \leq u)$ using $\{H_{i,j}(\lambda, \hat{\boldsymbol{N}}(\lambda))\}_{i=1,j=1}^{n,1000}$ for each $x = 1, \ldots, \mathcal{X}$
end
Set $\lambda_{\text{opt}} = \min \left\{ \lambda \in G_\lambda \text{ such that } \hat{P}_{\lambda,x} \geq p \text{ for all } x = 1, \ldots, \mathcal{X} \right\}$
Set $\boldsymbol{N}_{\text{opt}} = \hat{\boldsymbol{N}}(\lambda_{\text{opt}})$
return $(\lambda_{opt}, \boldsymbol{N}_{opt})$

the smallest λ that achieves the above inequality is selected. For example, besides providing on average 2 notifications per day, we might want to ensure that the probability of sending 1 to 3 notifications is at least $p = 0.95$ (e.g., $l = 1,\ u = 3$). In Appendix 4.5, we use a toy example to illustrate the selection of tuning parameters.

Recall that the forecast ($g(x|H_t)$) predicts the number of remaining available risk times at risk level x in the current time block. As pointed out earlier, the quality of the forecast determine the ability of SeqRTS in spreading out the treatments uniformly across the risk times at risk level x. Note that the size of time block also affects the forecast quality since the forecast needs to look more into the future if the size of the block is big. However, the block lengths should not be too short as then there will be blocks with no risk times. We suggest using a block length that is short,

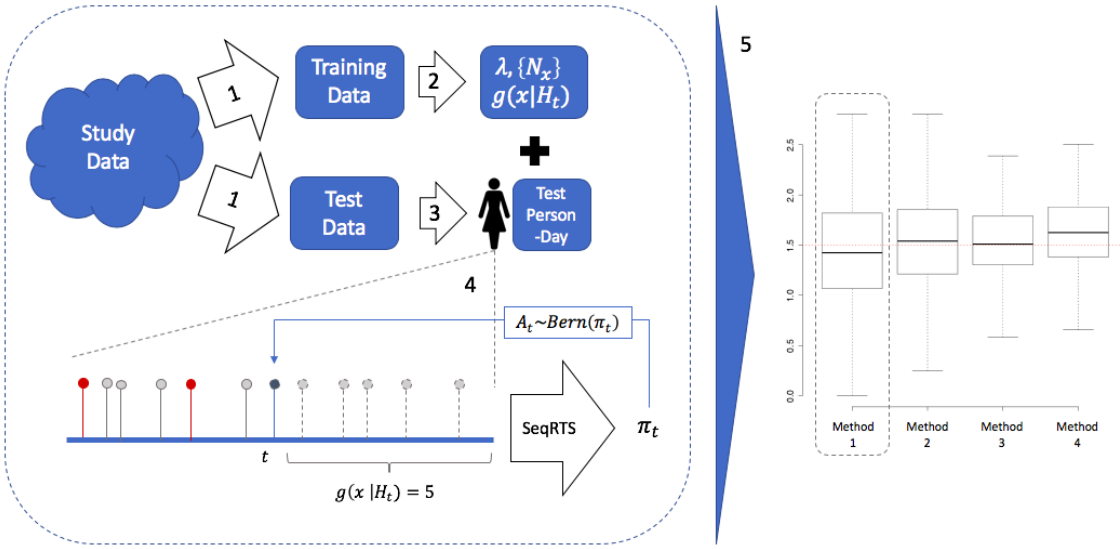

Figure 4.1: Flowchart for studying performance of the SeqRTS algorithm. In step 1, study data is split into training and test data. In step 2, TUNE is used to construct all tuning parameters and fit parameters for a chosen forecasting method. In Step 3, a person-day is extracted from the test data. In Step 4, SeqRTS is applied sequentially. The blue point indicates that time t is an available risk time at level x for the user; the prior red and gray points indicate past available risk times at level x. Prior points in red indicate treatment was provided. Combining this information with the forecast of 5 future available times at risk level x, SeqRTS is applied using the given tuning parameters and forecasting method to construct the probability of treatment at time t. In step 5, summaries of the performance on test data are aggregated using cross-validation. summaries for Method 1).

yet ensures that with high probability there will be at least N_x/K risk times at risk level x.

As the main focus of this work is on discussing the timing of treatment problem and the use of (4.2) along with how best to select the tuning parameters, we assumed that a method for forecasting is given. In practice, one can use the study data to both choose the tuning parameters and build the forecasts; we do this below. There are a number of existing methods for prediction and forecasting for the time-series, e.g., exponential smoothing and ARIMA model; see [58] for a review. In the Sense2stop example, we use forecasts obtained by Poisson regression.

Cross Validation Using Study Data The TUNE algorithm (2) selects tuning parameters using a training dataset. In both examples below, prior real mHealth studies exist that can be used for constructing training datasets and assessing performance via cross validation. Figure 4.1 visualizes the sequence of actions to perform cross validation and assess performance. In both examples, a fraction of the person-days from the study data is used as the training data, while the remainder is used as the test data (Step 1). The TUNE algorithm is applied to the training data to obtain

tuning parameters (i.e., Step 2). A particular person-day is extracted from the test data (i.e., Step 3). Then the SeqRTS algorithm is applied to test data (i.e., Step 4) to generate multiple treatment sequences; in our examples, 1000 treatment sequences are generated. Step 5 is cross validation to build performance summaries. This procedure outputs performance summaries for a particular chosen forecasting model. Therefore, the procedure must be run again for each proposed forecasting model. Alternatively, Step 1 could split the data by person rather than person-days; however, in these studies the resulting performance summaries are very similar.

4.3 Simulation

Here the SeqRTS algorithm is contrasted with a natural extension of the blocking method used in standard EMA setting. This extension considers the setting in which only risk times should be sampled for treatment and in which there is an average constraint on the number of treatments. This natural extension is a good comparator to our proposal because it is a simple extension that may achieve the desired soft constraint and uniformity across risk times.

To describe the extended version of block sampling method suppose there is only one level of risk (i.e., $X_t \in \{0, 1\}$) and the goal is to send treatment only at risk times (i.e., when $X_t = 1$). As in the standard blocking design, first construct K blocks of time within each day. If the average number of treatments per day is N, then in each block the goal is to provide an average of $n = N/K$ treatments. Suppose prior scientific knowledge and/or data from a prior study is used to estimate the number of expected risk times within each block, denoted by M_k for the k-th block. The number of blocks, K, would be chosen based on prior data and scientific rationale so that one can expect more risk times per block than the desired average number of treatments per block (i.e., $M_k > n$). Then at each risk time within each block, a treatment is sent with probability n/M_k. This "extended EMA blocking method" implicitly assumes that risk times are spread uniformly within blocks and that there is little between-user variability in the number of risk times per block.

Two metrics are used to compare SeqRTS with the above block sampling method. The first metric is graphical and the second metric uses a divergence function to assess divergence from uniform sampling of risk times for treatment. Recall that a training set is used to build the forecasts method as well as the select tuning parameters. These two metrics will be evaluated on a test data set. See Figure 4.1 for the cross validation procedure. Given the forecasts and selected tuning parameters, user-day trajectories within the test data are used to generate 1,000 treatment sequences (sequences of A_t's) per user-day. Each A_t is generated with the probability given by (4.2).

In the first, graphical metric, the 1000 generated treatment sequences are used to compute

the total number of treatments provided per day. The average over these treatment sequences is used to compute the average number of treatments per user-day. A box-plot across all user-day combinations of this average number of treatments summarizes performance. It is expected that the mean and median to agree with desired number of treatments per day, N_x^*. Methods with low variability around this mean are preferable. For the physical activity study, HeartSteps V1, for each user we compute the average of these user-day averages (i.e., average across days per user). A box-plot summarizes performance and highlights across-user variability in performance.

The second metric again uses the $1,000$ treatment sequences and computes the fraction of time treated per risk time at risk x for each day d. As treatment can only be provided at risk times, we extract these times out to construct the vector $\hat{p}_{u,d,x}$ (i.e., fraction of time treated for user u on day d at risk level x). Let $N_{u,d,x}$ be the number of risk times for user u on day d at risk x. Recall one of the goals for this algorithm is uniformity across risk times. To assess whether this goal is achieved we use the Kullback-Leibler divergence measure – a measure of "how one probability distribution diverges from a second, expected probability distribution" [59]. To assess uniformity, the second, "expected probability distribution" is the targeted uniform probability distribution across the true risk times. That is, knowing $N_{u,d,x}$ and the budget N_x, treatment should be provided marginally at each available risk time at level x with probability $N_x/N_{u,d,x}$. Therefore, the Kullback-Leibler divergence measured is between the sampling probabilities achieved by SeqRTS, e.g. $\hat{p}_{u,d,x}$ and the targeted uniform probabilities, $N_x/N_{u,d,x}$; that is,

$$\frac{1}{N_{u,d,x}} \sum_{i=1}^{N_{u,d,x}} (N_x^*/N_{u,d,x}) \log \left(\frac{\hat{p}_{u,d,x,i}}{N_x^*/N_{u,d,x}} \right)$$

where $\hat{p}_{u,d,x,i}$ is sampling probability for the ith risk time for user u on day d at risk level x vector $\hat{p}_{u,d,x}$. A box-plot of this quantity across user-days (all (u,d)'s) summarizes performance. Smaller Kullback-Leiber divergence indicates that the sampling is closer to uniform sampling and low overall variability indicates the sampling of risk times is closer to uniform for all user-days.

In the following, we first introduce the Minnesota smoking study which is used to design Sense2Stop, and evaluate and compare the performance of proposed algorithm, SeqRTS, with Extended Block Sampling method.

Minnesota Smoking Study

To design Sense2Stop, data from a smoking cessation study [35, 60] (here on called the "*Minnesota* smoking study") is used to construct forecasts, tune the parameters, and assess expected

performance. The Minnesota dataset is a no treatment, smoking study. Sensor data collected from wearable devices (e.g., the electrocardiogram (ECG) and respiration data) are used to produce an online, time-varying stress likelihood for each minute [34]. Next, a Moving Average Convergence Divergence approach (MACD) is adopted to locate an episode based on the time-series of stress likelihood; the episode consists of the start (the trend in stress likelihood changes from decreasing $(-)$ to increasing, $(+)$, the peak (from $+$ to $-$) and end time (from $-$ to $+$). At the peak time, a classification of the episode is made; we set $X_t = 2$ if the classification is "Stress", $X_t = 1$ if "Not Stress" and $X_t = 0$ is "Unknown"(when too much of the sensor data used for stress classification is missing or is of low quality due to sensor detachment or intermittent loosening); see [35] for details.

The dataset is restricted to user-days for which the duration of a day is at least 12 hours. Each user-day is truncated to 12 hours. This results in 54 user-days of 12 hours each. Three-fold cross validation is used to construct the training and test sets and assess performance. Specifically, user-days of the Minnesota dataset are randomly divided into three approximately equal subsets of user-day's data. Two subsets are used as the training set and the remaining one subset as the test set. This is repeated 2 further times to allow each subset to play the role of a test set. This reflects the procedure outlined in Figure 4.1.

Sense2Stop: Smoking Cessation Study

The performance assessment on the Minnesota smoking study informs expected performance in Sense2Stop [35, 50], an mHealth smoking cessation study currently underway; this study includes a 10-day post-smoking-quit phase in which users may receive a reminder to practice self-regulation exercises installed on the smart phone [38]. These exercises are designed to help users manage their stress as stress is a risk factor for relapse to smoking. Time frequency is every minute during a 12-hour day ($T = 720$). Sensor data collected from wearable devices matches the Minnesota smoking study (e.g., the same suite of sensors are worn), and the time-varying stress likelihood for each minute is computed in the same manner.

Availability I_t is set to 0 except for the peak time; furthermore, even at peak times, $I_t = 0$ if a treatment was provided within the prior hour or if self-report assessments (randomly assigned in each of 4-hour window) were requested from the user in the prior 10 minutes. Availability is similarly encoded in the study of the performance of the SeqRTS algorithm using the Minnesota study. Treatment at time t, A_t, is an indicator of whether a reminder is delivered at a time t (e.g., $1 =$ "deliver reminder" and $0 =$ "no reminder"). The goal is to provide, on average, 1 treatment at "Stress" and 1.5 treatments at "Not Stress" times per day, and no treatment if "Unknown;" that is,

$N_2^* = 1, N_1^* = 1.5$ and $N_0^* = 0$.

SeqRTS To design SeqRTS, the training set is used to build the forecasts for both the stress and non-stress episodes. There is a single block per day (i.e. $K = 1$) and thus, forecasts of the number of *available* stressed and non-stress episodes in the remaining of the day are required. Forecasts are built separately for stress and non-stress. Specifically, a Poisson regression is first fit using the training set with the outcome being the count of future "Stress" or "Not Stress" episodes and the input features: the remaining time of the day, the numbers of "Stress ", "Not Stress" and the "Unknown" episodes so far in the day and the indicator of lapse. To account for availability, these forecasts are further discounted by a constant (i.e. a guess of the fraction of future available stress/non-stress episodes). Bounds are set to $\epsilon_L = 0.01$ and $\epsilon_U = 0.99$. Next as described in subsection 4.2, the training data is used to select the tuning parameters λ and $\boldsymbol{N}$ using Algorithm 2—that is, for each λ, the parameters $\boldsymbol{N}$ are chosen to achieve the average constraints (an average of 1, 1.5 treatments per day at stressed, non-stress times respectively) and then the value of λ is chosen such that the probability of receiving at least 1 to at most 5 treatments (across both stress and non-stress episodes) in a day is at least 0.95.

Extended Block Sampling Here, the block sampling approach used in EMA setting to handle the risk setting with average constraints is adapted and used as a comparison to SeqRTS. As is customary with EMA, the 12-hour day is split into three four-hour blocks. Three-fold cross validation is applied as above using the procedure outlined in Figure 4.1. Using the training set, the average numbers of "Stress" episodes within each block in the training set are calculated and then discounted by a constant (to account for availability) to form an estimate of the number of available stress episodes in each block. Denote these numbers by M_1, M_2, M_3. This results in the block sampling method where, in each block, available stress times are randomly selected for treatment with probability $(1/3)/M_k$ in the k-th block at the "Stress" times (recall the goal is to send on average 1 treatment at "Stress" times). The same procedure is applied to "Not Stress" times.

Comparison of SeqRTS and Extended Block Sampling To compare SeqRTS with Extended Block Sampling, we use the test subset of the Minnesota data set. Since the Minnesota data set does not include all sources of un-availability, to more accurately represent availability as it occurs in Sense2Stop, we also generate the random self-report assessments (three times per day and randomly selected in each four-block block) and take into account availability constraints; here these are that the reminder messages can occur only after 10 or more minutes following a random self-report assessments and only after 60 or more minutes following a prior reminder message.

Using SeqRTS, we generate 1,000 treatment sequences for each real user-day in the test set. We also use Extended Block Sampling to generate 1,000 treatment sequences for each real user-day in the test set. For each user-day, we compute the average number of treatments at "Stress" and "Not stress" episodes and the percentage of sending 1 to 5 total treatments in a day over the 1,000 treatment sequences for both block sampling and the sequential sampling method. Recall we use 3-fold CV to train then test, thus the results are averaged over the 3 test sets. The results are shown in Figure 4.2 and Table 4.1. Recall that our goal is to achieve on average 1.5 reminders at "Not stress" and 1 reminder at "Stress" times and to ensure that with at least 0.95 probability of at least 1 and no more than 5 treatments are provided during the day. SeqRTS meets the desired desired average constraints; the average numbers of treatments at stress and not-stress over all user-days of 1.004 and 1.521. The block sampling method performs similarly in terms of the number of treatments at stress and not-stress times (0.912 and 1.426); however, SeqRTS is able to significantly reduce the variation of the average treatments across all 54 user-days in the test sets. For the "Stress" case, the standard deviations of the average treatments across user-days is 0.557 and 0.364 for extended block sampling and SeqRTS, respectively. For the "Not stress" case, the standard deviations are given by 0.612 and 0.298. The large variation of the average number of treatments for extended block sampling is due to the high variation of number of "Stress" and "Not Stress" times in each block, shown in Table 4.2. SeqRTS allows us to better achieve the average constraint across the user-days.

Table 4.1: Three-fold cross validation results of Extended Block Sampling (BS) and proposed Sequential Risk Times Sampling (SeqRTS): average number of treatments at "Stress" and "Not stress" times and the percentage of sending 1 to 5 total treatments achieved in each user-day across the 1,000 treatment sequences.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
# Treatments at NS (BS)	0.203	1.132	1.411	1.426	1.854	2.917
# Treatments at NS (SeqRTS)	0.275	1.454	1.601	1.521	1.713	1.830
# Treatments at S (BS)	0.172	0.451	0.766	0.912	1.317	2.225
# Treatments at S (SeqRTS)	0.245	0.797	1.037	1.004	1.240	1.705
Prob. of 1-5 Treatments (BS)	0.401	0.902	0.939	0.910	0.961	0.976
Prob. of 1-5 Treatments (SeqRTS)	0.785	0.929	0.947	0.948	0.958	0.991

Additionally extended block sampling produces high variance in the total number of treatments in a day (across the stress and non-stress times). The overall percentage of sending at least 1 and no more than 5 treatments in the blocking method is 0.910. SeqRTS controls this probability via the use of λ: the overall percentage is 0.948 as desired. The variations across the user-days is also significantly smaller (see Figure 4.2).

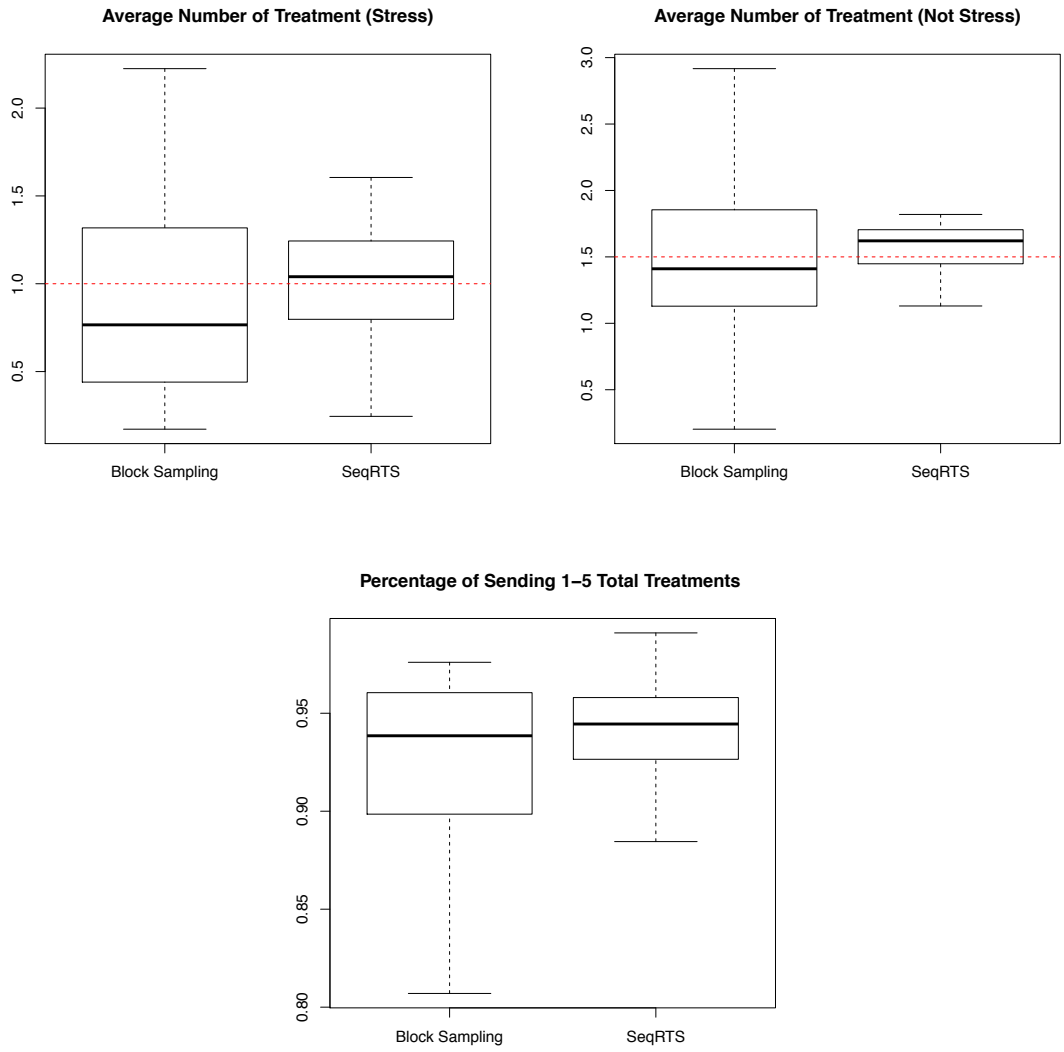

Figure 4.2: Three-fold cross validation results of Block Sampling and Sequential Risk Times Sampling (SeqRTS) algorithms. The average number of treatments at stress and not-stress episodes and the percentage of sending 1 to 5 total treatments achieved by each user-day in 1,000 runs.

Table 4.2: Summary statistics of the number of "Stress" and "Not Stress" times in each block. MAD: mean absolute deviation.

	Not Stress			Stress		
	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3
Mean	8.98	8.37	7.52	1.65	1.59	1.74
MAD	3.91	4.35	3.48	1.51	1.30	1.65

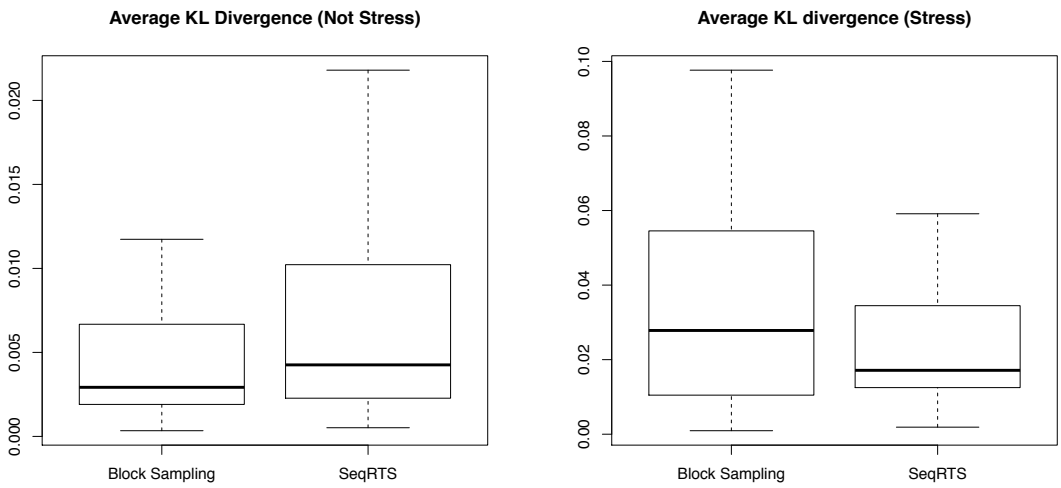

Figure 4.3: Three-fold cross validation results of Block Sampling and Sequential Risk Times Sampling (SeqRTS) algorithms. The average KL divergence at stress and not-stress episodes by each user-day in 1,000 runs.

As discussed in the beginning of this section, we assess whether a method achieves uniform sampling across risk times via the average KL divergence. The results are provided in Figure 4.3. We see that SeqRTS achieves smaller average KL over all user-days in the case of "Stress" times comparing with extended block sampling (i.e. more uniform distribution of treatment across "Stress" times), but larger in the "Not Stress" times. The latter is mainly due to the quality of the estimated forecast. This can be seen as follows. We run SeqRTS replacing the forecast of future stress and non-stress episodes by the weighted average of the estimated forecast from the model and the oracle (i.e. the true number of the future "Stress" and "Not Stress" episodes). As we can see in Figure 4.4, the average KL can be significantly reduced comparing with extended block sampling when we have high quality forecasts. The results of average number of treatments and variation in the total treatments are similar using these hypothetical forecasts and thus are omitted.

4.4 Conclusion and Future Work

In this chapter, we have proposed a new approach to design the timing of just-in-time treatments when there is a soft constraint on the number of treatments per day. We have illustrated how one selects tuning parameters so as to achieve the constraint, yet maintain an acceptable level of variance in number of treatments delivered. If there is within- and between-user variation in

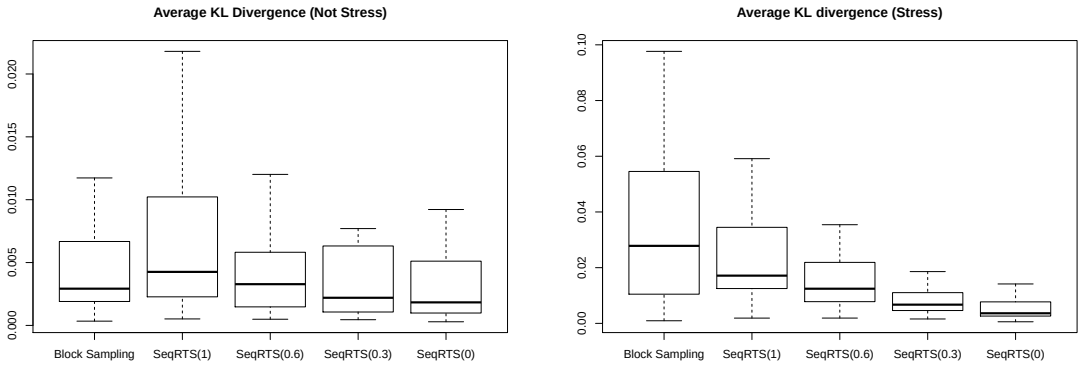

Figure 4.4: Three-fold cross validation results of Block Sampling (BS) and Sequential Risk Times Sampling algorithms (SeqRTS(w)) method using hypothetical forecasts : the average KL divergence at "Stress" and "Not Stress" episodes by each user-day in 1,000 runs. w is the proportion of the estimated forecasts in the constructing the hypothetical forecasts used in SeqRTS

the risk pattern and this variability is well explained by (time-varying) covariates, then the use of forecasts based on these time-varying covariates allows SeqRTS to sample risk times with a uniform distribution thus providing data that allows scientists to learn if and in which contexts the treatment is effective. Extended block sampling achieves uniform sampling of the risk times when there is minimal within- and between-user variation in the risk pattern.

We foresee several opportunities for future work. First, viewing SeqRTS as a warm-start, one could update the tuning parameters and the forecasts as information accumulates on a user during the study. This personalization might occur by making the tuning parameters person-specific. Second, depending on the amount of training data, a variety of forecasting algorithms might be considered including deep learning algorithms, such as the Long Short Term Memory algorithm; these methods would facilitate the investigation of how to combine/fuse multiple data streams (stress, location, eating, etc.) in forecasts and thus enable prediction of more complex risk variables. Finally, there it would be of interesting to develop a version of the SeqRST algorithm that includes assumptions about non-zero treatment effects.

4.5 Appendix

A. Derivation of Randomization Probabilities

Here we provide a brief discussion of how the randomization probability formulation $\pi_t = \pi_t(H_t)$ in (4.2) is motivated and derived. For simplicity, below we consider deriving the probability to

satisify the average budget constraint for a single time block with T decision time point, e.g., $\mathbb{E}\big[\sum_{t=1}^{T} A_t \mathbb{1}_{\{X_t=x, I_t=1\}}\big] = N_x^*$. We start by re-writing LHS of equation: for arbitrary $\lambda_t \in [0,1]$,

$$\mathbb{E}\big[\sum_{t=1}^{T} A_t \mathbb{1}_{\{X_t=x, I_t=1\}}\big] = \mathbb{E}\big[\sum_{t=1}^{T} \big(\lambda_t A_t + (1-\lambda_t)\pi_t(H_t)\big)\mathbb{1}_{\{X_t=x, I_t=1\}}\big]$$

A natural goal here is to find $\pi_s = \pi_s(H_s)$ for $s \in \{1, \ldots, T\}$ such that for each $x \in [k]$,

$$\begin{aligned} N_x^* &\approx \mathbb{E}\big[\sum_{t=1}^{T} \big(\lambda_t A_t + (1-\lambda_t)\pi_t(H_t)\big)\mathbb{1}_{\{X_t=x, I_t=1\}} \,|\, H_s\big] \\ &= \sum_{t=1}^{s-1} \left(\lambda_t A_t + (1-\lambda_t)\pi_t(H_t)\right)\mathbb{1}_{\{X_t=x, I_t=1\}} + \pi_s(H_s) + \mathbb{E}\big[\sum_{t=s+1}^{T} \pi_t(H_t)\mathbb{1}_{\{X_t=x, I_t=1\}} \,|\, H_s\big] \end{aligned}$$

The first term is known given the current history H_s. On the other hand, at the decision time s, we do not have access to future randomization probabilities, i.e. $\pi_t(H_t)$ under $X_t = x$ and $I_t = 1$ for $t \geq s+1$, which appears in the last term above. As such, we pretend $\pi_t(H_t) = \pi_s(H_s)$ whenever $I_t = 1$, that is, we use the exact same randomization probabilities for future time available risk points, and obtain $\pi_s = \pi_s(H_s)$ by solving:

$$N_x = \sum_{t=1}^{s-1} \left(\lambda_t A_t + (1-\lambda_t)\pi_t(H_t)\right)\mathbb{1}_{\{X_t=x, I_t=1\}} + \pi_s(H_s) + \mathbb{E}\big[\sum_{t=s+1}^{T} \pi_s(H_s)\mathbb{1}_{\{X_t=x, I_t=1\}} \,|\, H_s\big]$$

This implies

$$\pi_s(H_s) = \frac{N_x - \sum_{t=1}^{s-1}\left[\lambda_t A_t + (1-\lambda_t)\pi_t(H_t)\right]\mathbb{1}_{\{X_t=x, I_t=1\}}}{1 + \mathbb{E}\big[\sum_{t=s+1}^{T} \mathbb{1}_{\{X_t=x, I_t=1\}} \,|\, H_s\big]}$$

Note that in the denominator $\mathbb{E}[\sum_{t=s+1}^{T} \mathbb{1}_{\{X_t=x, I_t=1\}} \,|\, H_s]$ is the forecast of the number of available risk times at risk level x. By choosing $\lambda_t = \lambda^{s-t}$ and restricting the probability within $[\epsilon_L, \epsilon_U]$, we obtain the randomization probability formula (4.2).

B. Toy Example

Below we use a toy example to illustrate how the tuning parameters $(\boldsymbol{N}, \lambda)$ and the forecasts impact the performance of the SeqRTS algorithm and illustrate the selection of tuning parameters. For simplicity, consider the case where the user is always available for treatment (i.e., $I_t = 1$) and the risk variable is binary, $X_t \in \{0, 1\}$. Temporal frequency is every fifteen minutes and we consider providing treatments in a 10-hour day ($T = 40$). In this toy example, there is a single

level of risk (i.e., $X_t = 1$) and a single whole day block for simplicity (i.e. $K = 1$). The risk variables $\{X_1, \ldots, X_T\}$ are generated i.i.d. with probability 0.5. The goal is to provide on average 3 treatments per day (i.e., $N_0^* = 0$ and $N_1^* = 3$) at risk times. The training dataset is 100 randomly generated user-days.

A series of three simulations illustrate the performance of the SeqRTS. The first simulation (S1) illustrates how the tuning parameters impact the average number of daily treatments. Algorithm performance is evaluated with tuning parameters $N_{1,1} \in \{2.95, 3, 3.05, 3.1\} := G_{\boldsymbol{N}}$ and $\lambda \in \{0, 0.1, \ldots, 0.9\} := G_\lambda$ using the test set. In the second simulation (S2), the value of $N_{1,1}$ is tuned (using training set) so as to achieve the average constraint under each λ. In both S1 and S2, the forecasts are correct in expectation, e.g. $g(1|H_t) = 0.5(T - t)$.

The simulation results of both (S1) and (S2) are provided in Figure 4.5. The left graph in this figure illustrates $N_{1,1}$ together with the choice of λ control the total number of notifications. The appropriate value of $N_{1,1}$ to achieve the average constraint (i.e., 3 in this toy example) depends on the value of λ. For example, when $\lambda = 0$, $N_{1,1}$ needs to be greater than 3, whereas in the case of $\lambda = 0.8$, the appropriate choice of $N_{1,1}$ is less than 3. After properly choosing $N_{1,1}$ for each λ, the right graph in Figure 4.5 shows that incorporating λ allows the algorithm to control the variability in number of treatments yet achieve the average constraint. The y-axis is the probability that the number of treatments sent lies between 1 and 5.

The last simulation (S3) illustrates the impact of an inaccurate forecasting method on SeqRTS. A class of forecasts indexed by a constant τ are considered, i.e., $g(1|H_t) = (T - t)\tau$. For each forecast, the tuning parameter λ is chosen using training set to be the smallest one over the grid set $\{0, 0.05, \cdots, 0.95\}$, such that with at least 0.95 probability the number of treatments sent lies in the range of 1 to 5. Here, the parameter $N_{1,1}$ is tuned for each λ as in S2. In all cases, $N_{1,1}$ can be tuned to achieve the average constraint of 3, and λ tuned to achieve at least 0.95 probability that the number of treatments sent lies in the range of 1 to 5. However, as shown in Figure 4.5, the more inaccurate the forecast (i.e. τ far from 0.5), the less uniform the distribution of treatments assigned across hour blocks (note the times when $X_t = 1$ are uniformly distributed across time in this toy example).

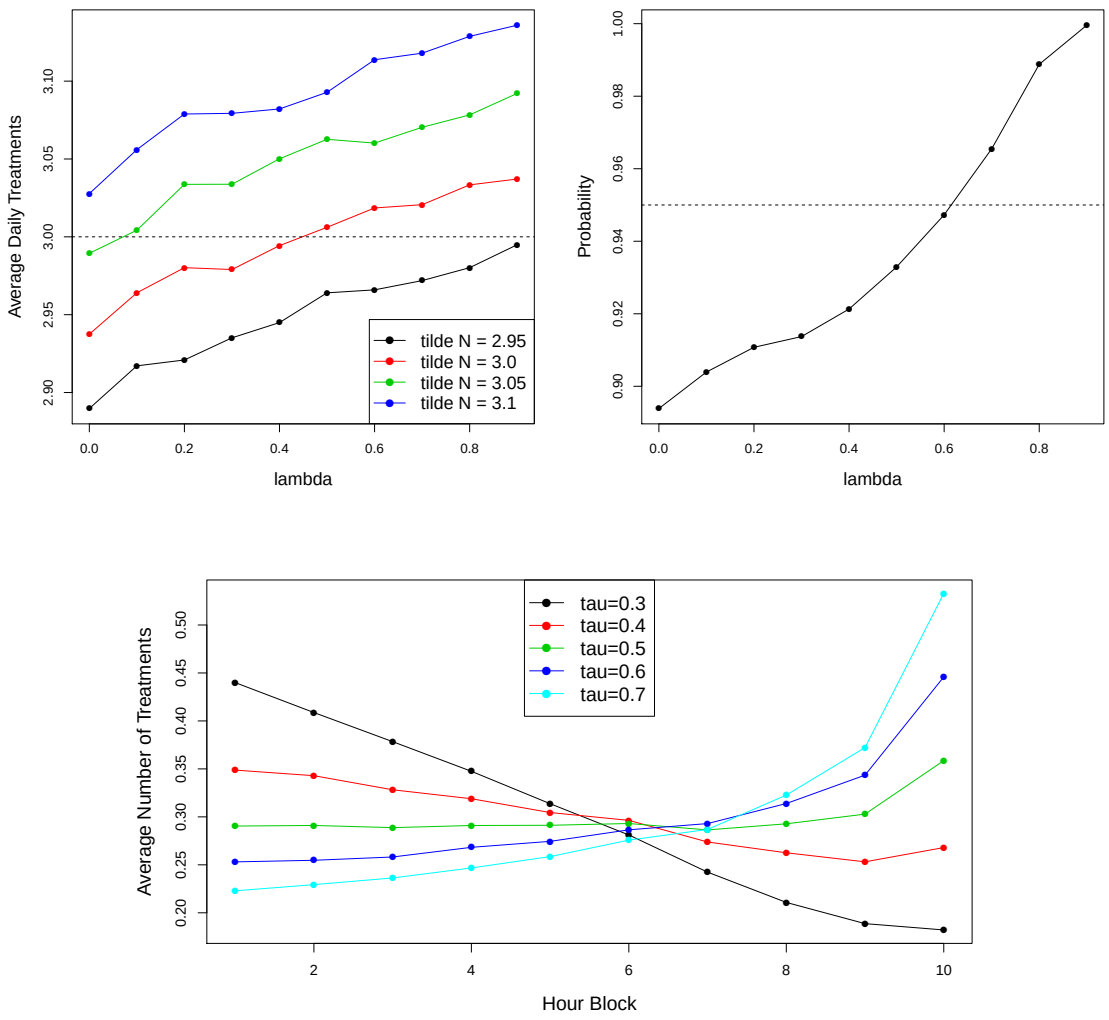

Figure 4.5: Simulation results of toy example. *Top Left*: the average number of daily treatments under different values of λ (x-axis) and $N_{1,1}$ (in color) in (S1). *Top Right*: the probability that the number of treatments sent ranges between 1-5 treatments in (S2) Here $N_{1,1}$ is tuned for each λ. Note, from the range of the y-axis that the probability ranges from 0.89 to 1.00 in this example. *Bottom*: Average number of treatments triggered in each hour block in testing data set in (S3). Solid lines = "forecast"; dashed line = "oracle"

CHAPTER 5

Inference of Long-Term Average Outcome

5.1 Introduction

With the recent evolution of mobile health technologies, health scientists are increasingly interested in developing Just-In-Time Adaptive Interventions (JITAIs), typically delivered via a notification on the mobile device and designed to help the user prevent negative health outcomes and promote the adoption and maintenance of healthy behaviors. JITAIs can be operationalized by a sequence of decision rules (e.g., treatment policies) that takes the users current context as input and specifies whether and what type of an intervention should be provided at the moment.

However, the vast majority of current deployed JITAIs are theory-guided; the decision rules underpinning JITAIs are formulated using domain expertise and clinical experience, with very limited use of data evidence. This approach is unlikely to achieve the full potential of JITAIs, since most of the current behavior theories and empirical evidence fail to specify the dynamics of human behavior at a level that is sufficient enough to be able to personalize the interventions in the best way possible, and more importantly, to optimize the long-term efficacy of the interventions as a whole.

In this chapter, we take a step toward developing a data-based approach to inform the construction of efficacious JITAIs. In particular, we develop a batch data analysis method for estimating the average of the long-term positive health outcomes (i.e., reward) that would accrue should a given JITAI (i.e., the target policy) be followed. Furthermore, we develop the inferential procedure to construct the confidence intervals for the estimates. The method can also be used to contrast multiple JITAIs by comparing the long-term average outcomes. The method uses a training data, collected under a possibly different policy, called behavioral policy. This is the setting where the data is collected from Micro-randomized trials in which the treatments are randomly selected and is different from the treatment policy of scientific interest. In Reinforcement Learning (RL) literature this is also called the off-policy evaluation problem.

The proposed method is developed in the Markov Decision Process (MDP), a common framework used in RL. We use the average reward criteria in the infinite-horizon setting. Another commonly used criterion in the infinite horizon is the discounted reward setting. Most of RL literature consider the discounted reward setting. There are mainly two different types of approaches: model-free and model-based approach. In the latter, the estimate of the value function is done by first building a system dynamic model. In the model-free method, which is the method we consider in this work, a function class is deployed to estimate the value function based on the so-called Bellman equation. One of the first approaches in the model-free, batch policy evaluation problem is called Least Square Temporal Difference (LSTD), first published in [61]. In the original LSTD, a linear model is used to approximate the value function. There are many analysis for LSTD, for example in [62, 63, 64, 65] and many variants of LSTD in the literature. More recently, in [66] a regularized version of LSTD was proposed and the statistical property was studied. More specifically, they used a non-parametric model to estimate the value function and derive the convergence rate when training data consists of i.i.d. samples. Our proposed approach is similar to [66], but we extend it to the multiple trajectories in average reward setting and relax one of the key assumptions in their method. More importantly, we develop the semi-parametric inference of the estimated average reward. Another closely related work is [67], in which they developed "V-learning" in the discounted reward setting for off-policy learning in mobile health. For each target policy, they proposed to learn the average of the discounted value and developed a regularized estimating equation to estimate the value function using a parametric model. In contrast, here we consider the long-term average reward of the target policy and develop a semi-parametric method where the relative value function is viewed as the nuisance parameter.

In the following section, we introduce Markov Decision Process and the average reward criterion. In Section 5.3 we introduced the proposed estimator. The main theoretical results are presented in Section 5.4. We also develop three important generalizations of the proposed method in Section 5.5. We end with a discussion of future work in Section 5.6.

5.2 Markov Decision Process

We model the sequential decision making process as a Markov Decision Process (MDP). Consider

$$\{S_1, A_1, S_2, A_2, S_3, \dots, S_t, A_t, S_{t+1}, \dots\},$$

where t indexes the decision time, $S_t \in \mathcal{S}$ is the state variable and $A_t \in \mathcal{A}$ is the action selected at time t. We assume the action space is finite and the data-generating process is Markovian, e.g., for $t \geq 1$, $S_{t+1} \perp \{S_1, A_1, \ldots, S_{t-1}, A_{t-1}\} \,|\, \{S_t, A_t\}$, and time-invariant. Let $\mathcal{P}(\mathcal{X})$ denote the class of distribution on $\mathcal{X}$. Denote the transition kernel by $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ so that $P(\cdot|s,a)$ is the distribution of next state given the current state-action pair (s,a). Denote by $p(s'|s,a)$ the transition density with respect to some reference measure on $\mathcal{S}$ (e.g., counting measure when $\mathcal{S}$ is discrete). The reward is defined as a known function of the tuple (S_t, A_t, S_{t+1}) at each time t and denoted by $R_{t+1} = \mathcal{R}(S_t, A_t, S_{t+1})$. We also use $r(s,a)$ to denote the conditional expectation of reward given state and action, i.e., $r(s,a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$.

A policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a mapping that takes the state as input and outputs a probability distribution on the action space $\mathcal{A}$. Let $\pi(a|s)$ be the probability of selecting action a given state s. In this paper, we evaluate the policy using the long-term average reward. Specifically, the average reward of a policy π is defined as

$$\eta^\pi(s) = \lim_{T\to\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{R}(S_t, A_t, S_{t+1}) \,\Big|\, S_1 = s \right] \tag{5.1}$$

where the expectation, $\mathbb{E}_\pi$, is taken over the trajectory $\{S_1, A_1, S_2, \ldots, S_T, A_T, S_{T+1}\}$ in which the actions $\{A_t\}_{t\geq 1}$ are selected according to the policy π with the starting state $S_1 = s$, that is, the likelihood in the expectation is given by $\mathbb{1}_{\{S_1=s\}} \prod_{t=1}^{T} \pi(A_t|S_t) p(S_{t+1} \,|\, S_t, A_t)$. Note that the policy π induces a Markov Chain with transition kernel $P^\pi(\cdot|s) = \sum_a \pi(a|s) P(\cdot|s,a)$.

Suppose for now the state space $\mathcal{S}$ is finite. It is well known [68] that when the induced markov chain P^π is irreducible and aperiodic, the average reward defined in (5.1) is independent of the initial state, i.e.,

$$\eta^\pi(s) = \eta^\pi = \int_{\mathcal{S}} \sum_a \pi(a|s) r(s,a) d^\pi(s) ds \tag{5.2}$$

where $d^\pi(s)$ is the density of the stationary distribution (the existence is guaranteed by irreducibility and aperiodicity). Furthermore, we can define the relative value function Q^π

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} (\mathcal{R}(S_t, A_t, S_{t+1}) - \eta^\pi) \,\Big|\, S_1 = s, A_1 = a \right] \tag{5.3}$$

It is easy to verify by definition that (η^π, Q^π) is the solution of Bellman (Policy Evaluation) equa-

tion (also known as Poisson equation), given as follows: for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

$$\mathbb{E}_\pi[R_{t+1} + Q(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a] = \eta + Q(s, a). \tag{5.4}$$

Bellman equation uniquely identifies the average reward and identifies the value function Q^π up to a constant. That is, the set of solution of (5.4) is given by $\{(\eta^\pi, Q) : Q = Q^\pi + c\mathbf{1}, c \in \mathbb{R}, \mathbf{1}(s, a) = 1\}$. See [68] for details. The above results can be generalized to general state space, e.g., $\mathcal{S} \subset \mathbb{R}^d$, with more involved conditions on the transition kernel P^π, analogous to irreducibility and aperiodic in the finite state case; see Chapter 7 in [69]. The key requirement for the proposed method presented in Section 5.3 is that the average reward is a constant and can be uniquely identified by solving Bellman equation (5.4). We will consider the generalization that allows the average reward to depend on a time-invariant state in Section 5.5.

5.3 Off-Policy Evaluation

We consider the setting where we have access to a training data with n independent, identically distributed trajectories:

$$\mathcal{D}_n = \{S_1^i, A_1^i, S_2^i, \ldots, S_T^i, A_T^i, S_{T+1}^i\}_{i=1}^n.$$

We assume the length of trajectory, T, is non-random and identical for each trajectory for simplicity. Each trajectory $\mathcal{D} = \{S_1, A_1, S_2, \ldots, S_{T+1}\}$ is assumed to follow a MDP in which the actions selected according some behavioral policy π_b, i.e., $A_t \sim \pi_b(\cdot \,|\, H_t)$ where $H_t = \{S_1, A_1, \ldots, S_t\}$ is the history collected up to decision time t. In what follows, the expectation $\mathbb{E}$ without the subscript is assumed taken with respect to the distribution of the trajectory $\mathcal{D}$ with the actions selected by the behavioral policy π_b. Let ν_t be the marginal distribution of the state-action pair $\{S_t, A_t\}$ in the training data (e.g., the previous actions are selected according to behavior policy π_b) and let $\bar{\nu}_T$ be the average distribution across T decision times, e.g., $\bar{\nu}_T = (1/T)\sum_{t=1}^T \nu_t$. Denote by d_t and $\bar{d}_T$ the density (mass) function of ν_t and $\bar{\nu}_T$. For any function of state and action $f(s, a)$ and distribution ν, denote the L_2 norm by $\|f\|_\nu^2 = \int f^2(s, a) d\nu(s, a)$. Note that we have $\|f\|_{\bar{\nu}_T}^2 = \mathbb{E}[(1/T)\sum_{t=1}^T f^2(S_t, A_t)]$.

In mobile health applications, the action A_t is the treatment or intervention option and the states S_t contains the time-varying contextual information (e.g. stress, location, busyness in the calendar) and summary of historical data up to and including time t (e.g. summary of previous physical activity). Let π be some target policy, possibly different from the behavior policy π_b in the training data. Throughout we only consider the target policy that is Markovian (i.e., only

depends on the current state) and time-invariant (i.e. the mapping does not vary with time). Our goal is to learn η^π, the average reward of the target policy. We assume the target policy satisfies the following.

Assumption 1. *The average reward of the target policy π is independent of state and satisfies (5.2) and (η^π, Q^π) is the unique solution of Bellman equation (5.4) up to a constant for Q^π. The stationary distribution of induced transition kernel P^π exists and the density is denoted by $d^\pi(s)$.*

Note that the Assumption 1 prevents us from including time-invariant states, e.g., participant's demographic information. We will consider this extension in Section 5.5. We consider a model-free approach to estimate the average reward based on Bellman equations (5.4). Recall that Under Assumption 1, Bellman equations can only identify the value function up to a constant. As the focus of this paper is to estimate the average reward, we only need to estimate one specific version of value function. We define the shifted value function. Given a specific state-action pair (s^*, a^*), denote by $\tilde{Q}^\pi(s,a) = Q^\pi(s,a) - Q^\pi(s^*,a^*)$ the shifted value function. Obviously the shifted value function $\tilde{Q}^\pi(s^*,a^*) = 0$ and $\tilde{Q}^\pi(s_1,a_1) - \tilde{Q}^\pi(s_2,a_2) = Q^\pi(s_1,a_1) - Q^\pi(s_2,a_2)$, e.g., the difference in the value remains the same. By restricting the function class s.t., $Q(s^*,a^*) = 0$, the solution of Bellman equations (5.4) is unique and given by $(\eta^\pi, \tilde{Q}^\pi)$.

In the following, we use $\mathcal{Q}$ to denote a vector space of functions on the state-action space $\mathcal{S} \times \mathcal{A}$ such that $Q(s^*,a^*) = 0$ for all $Q \in \mathcal{Q}$, in which we will assume $\tilde{Q}^\pi \in \mathcal{Q}$. Motivated from the Bellman equation, we introduce the Bellman error operator $\mathcal{E}$ with respect to the target policy π. For any $(\eta, Q) \in \mathbb{R} \times \mathcal{Q}$, define the temporal difference error $\delta(S, A, S', R; \eta, Q) = R + \sum_{a'} \pi(a'|S')Q(S',a') - \eta - Q(S,A)$ and

$$\mathcal{E}(s,a;\eta,Q) = \mathbb{E}[\delta(S_t, A_t, S_{t+1}, R_{t+1}; \eta, Q) \mid S_t = s, A_t = a] \tag{5.5}$$

Note that the Bellman error is not necessarily in the function space $\mathbb{R} \oplus \mathcal{Q}$. To see this, plugging $\mathbb{E}[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})\tilde{Q}^\pi(S_{t+1},a') - \eta^\pi - \tilde{Q}^\pi(S_t,A_t)|S_t = s, A_t = a] = 0$ implies that

$$\begin{aligned}\mathcal{E}(s,a;\eta,Q) =& \mathbb{E}[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1},a') - \eta - Q(s,a) \mid S_t = s, A_t = a] \\ =&(\eta^\pi - \eta) + (\tilde{Q}^\pi - Q)(s,a) - \mathbb{E}[\sum_{a'} \pi(a'|S_{t+1})(\tilde{Q}^\pi - Q)(S_{t+1},a')|S_t = s, A_t = a]\end{aligned}$$

Depending on the complexity of the transition kernel, the last term unlikely stays in $\mathcal{Q}$ for every $Q \in \mathcal{Q}$. Because of this, we introduce another linear function class $\mathcal{G}$ to form a surrogate Bellman

error operator $\mathcal{E}_{\mathcal{G}}$, defined as

$$\mathcal{E}_{\mathcal{G}} : (\eta, Q) \to \mathcal{E}_{\mathcal{G}}(\cdot, \cdot; \eta, Q) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathbb{E}\big[(1/T) \sum_{t=1}^{T} (\mathcal{E}(S_t, A_t; \eta, Q) - g(S_t, A_t))^2\big] \tag{5.6}$$

We now construct the partially penalized estimator for $(\eta^\pi, \tilde{Q}^\pi)$. In what follows, for ease of notation we use $\hat{Q}_n$ to denote the estimates of $\tilde{Q}^\pi$, the shifted value function. For an arbitrary function f, let $\mathbb{P}_n f(H_t) = (1/n) \sum_{i=1}^{n} f(H_t^i)$ be the empirical mean over the training data $\mathcal{D}_n$. The partially penalized estimator is found by minimizing the mean squared estimated Bellman error plus a penalty term on the value function:

$$(\hat{\eta}_n, \hat{Q}_n) = \underset{(\eta, Q) \in \mathbb{R} \times \mathcal{Q}}{\operatorname{argmin}} \mathbb{P}_n \left[\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{E}}_n^2(S_t, A_t; \eta, Q) \right] + \lambda_n J_1^2(Q) \tag{5.7}$$

where $J_1 : \mathcal{Q} \to \mathbb{R}^+$ is the regularizer, λ_n is a tuning parameter and $\hat{\mathcal{E}}_n(\cdot, \cdot; \eta, Q)$ is an estimator of the surrogate Bellman error operator $\mathcal{E}_{\mathcal{G}}(\cdot, \cdot; \eta, Q)$ defined in (5.6). Specifically, for all $(\eta, Q) \in \mathbb{R} \times \mathcal{Q}$, $\hat{\mathcal{E}}_n(\cdot, \cdot; \eta, Q) \in \mathcal{G}$ is a function of state-action pair in the class $\mathcal{G}$ that minimizes the least squared error with a penalty:

$$\hat{\mathcal{E}}_n(\cdot, \cdot; \eta, Q) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathbb{P}_n \left[\frac{1}{T} \sum_{t=1}^{T} \big(\delta_t(\eta, Q) - g(S_t, A_t)\big)^2 \right] + \mu_n J_2^2(g) \tag{5.8}$$

where $\delta_t(\eta, Q) = \delta(S_t, A_t, S_{t+1}, R_{t+1}; \eta, Q)$. Similar to J_1 and λ_n, $J_2 : \mathcal{G} \to \mathbb{R}^+$ is a regularizer on the function space $\mathcal{G}$ and μ_n is tuning parameter. The penalty term $\lambda_n J_1^2(Q)$ is used to balance between the model fitting, i.e. the squared estimated Bellman error and the complexity of the value function, measured by $J_1(Q)$. Similarly, $\mu_n J_2^2(g)$ is used to control the overfitting in estimating the Bellman error when the space $\mathcal{G}$ is complex. In the case where the function space is k-th order Sobolev space, the regularizer is typically defined by the k-th order derivative to capture the smoothness of function. In the case where the function space is Reproducing Kernel Hilbert Space (RKHS), the regularizer is the endowed norm. In Appendix 5.7, we provide a closed-form solution of the estimator when both $\mathcal{Q}$ and $\mathcal{G}$ are RKHSs.

Linear Approximation and L_2 Regularization In what follows we explain the estimator when using linear approximations. Consider a feature vector $\phi(s, a) \in \mathbb{R}^p$ such that $q(s^*, a^*) = 0$ and $g(s, a) \in \mathbb{R}^q$. Define $\mathcal{Q} = \{\phi(\cdot, \cdot)^\top q : q \in \mathbb{R}^p\}$ and $\mathcal{G} = \{g(\cdot, \cdot)^\top \theta : \theta \in \mathbb{R}^q\}$. Consider the standard L_2 regularization (e.g. the ridge penalty), $\|q\|_2^2$ and $\|\theta\|_2^2$. In this case, the estimator (5.7)

can be found by

$$(\hat{\eta}_n, \hat{q}_n) = \text{argmin}_{\eta,q} \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \delta_t(\eta, q) g(S_t, A_t)\big]^\top M \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \delta_t(\eta, q) g(S_t, A_t)\big] + \lambda_n \|q\|_2^2$$

where $M = (\hat{\Sigma}_n + \mu_n I_q)^{-1} \hat{\Sigma}_n (\hat{\Sigma}_n + \mu_n I_q)^{-1}$ and $\hat{\Sigma}_n = \mathbb{P}_n[(1/T)\sum_{t=1}^{T} g(S_t, A_t) g(S_t, A_t)^\top]$. We see that the estimator can be viewed as regularized estimating equation that solves

$$\mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T} \delta_t(\eta, q) g(S_t, A_t)\Big] = 0_{q+1}$$

In [67], they consider the in the discounted reward setting and developed a similar regularized estimating equation with a non-random matrix M. In addition, they construct the feature vector $g(\cdot, \cdot)$ by taking the derivative of the value function. Transferring into the average reward setting, this implies the choice of $g(s, a) = (1, \phi(s, a))^\top$.

5.4 Theoretical Results

In this section, we derive the global rate of convergence for the $(\hat{\eta}_n, \hat{Q}_n)$ in (5.7) and derive the asymptotic distribution of $\hat{\eta}_n$. We make the following assumptions.

Assumption 2. *The reward function is bounded:* $\mathcal{R}(s, a, s') \leq R_{\max}$ *for all* (s, a, s') *tuple. The shifted value function is bounded:* $|\tilde{Q}^\pi(s)| \leq Q_{\max}$ *for all* $s \in \mathcal{S}$.

The bounded reward is mainly to simplify the proof and can be easily relaxed to the sub-Gaussian case, e.g. the error $R_{t+1} - r(S_t, A_t)$ is sub-Gaussian. The boundedness assumption of the value function can be ensured by assuming certain smoothness assumption on the transition distribution [70] or assuming geometric convergence to the stationary distribution (see [69]).

Assumption 3. *The function class* $\mathcal{Q}$ *satisfies that*

(i) $Q(s^*, a^*) = 0$ *and* $\|Q\|_\infty \leq Q_{\max}$ *for all* $Q \in \mathcal{Q}$,

(ii) $\tilde{Q}^\pi \in \mathcal{Q}$, *where* $\tilde{Q}^\pi = Q^\pi - Q^\pi(s^*, a^*)$ *is the shifted value function.*

Assumption 4. *The function class* $\mathcal{G}$ *satisfies that*

(i) $\|g\|_\infty \leq G_{\max}$ *and*

(ii) $\kappa = \inf\big\{\|\mathcal{E}_{\mathcal{G}}(\cdot, \cdot; \eta, Q)\|_{\bar{\nu}_T} : \|\mathcal{E}(\cdot, \cdot; \eta, Q)\|_{\bar{\nu}_T} = 1, \eta \in \mathbb{R}, Q \in \mathcal{Q}\big\} > 0$

The correct modeling of $\tilde{Q}^\pi$ in (ii) of Assumption 3 is reasonable when $\mathcal{Q}$ is rich enough. The result in Theorem 1 can be generalized to allow the approximation error e.g., when $\tilde{Q}^\pi \notin \mathcal{Q}$. However the asymptotic inference of the average reward requires the correct modeling of the value function and thus we assume it throughout for simplicity. The boundedness assumption of $\mathcal{Q}$ is only used to simplify the proof. A truncation argument can be used to avoid this assumption. Recall that for a state-action function $f(s,a)$, the norm $\|f\|_{\bar{\nu}_T}^2 = \mathbb{E}[(1/T)\sum_{t=1}^T f^2(S_t, A_t)]$. The value of $\kappa \in [0,1]$ measures the quality of how well the function class $\mathcal{G}$ approximates the Bellman error for all (η, Q) in which $\eta \in \mathbb{R}$ and $Q \in \mathcal{Q}$. A strictly positive value of κ ensures we have a consistent estimate $(\eta^\pi, \tilde{\mathcal{Q}}^\pi)$. Note that unlike in [66], here we do not assume $\mathcal{E}(\cdot,\cdot;\eta,Q)$ is modeled correctly by $\mathcal{G}$ for every (η, Q). In fact, in this case we have $\kappa = 1$. Note that $\mathcal{E}_{\mathcal{G}}(\eta^\pi, \tilde{Q}^\pi) = 0$. The condition of a strict positive value of κ ensures the estimator (5.7) based on minimizing projected Bellman error onto the space $\mathcal{G}$ is able to identify the true parameters $(\eta^\pi, \tilde{Q}^\pi)$. This is similar to the eigenvalue condition (Assumption 5) in [67].

Assumption 5. *(i) The regularization functional J_1 and J_2 are pseudo norms and induced by the inner products $J_1(\cdot,\cdot)$ and $J_2(\cdot,\cdot)$, respectively. (ii) For all $(\eta, Q) \in \mathbb{R} \times Q$, there exists two constants C_1, C_2 such that $J_2(\mathcal{E}_{\mathcal{G}}(\cdot,\cdot;\eta,Q)) \le C_1 + C_2 J_1(Q)$*

The inner product is sufficient to derive the asymptotic distribution of $\sqrt{n}(\hat{\eta}_n - \eta^\pi)$. This is satisfied for most common function class, for example RKHS and Sobolev space. The upper bound of $J_2(\mathcal{E}_{\mathcal{G}}(\eta, Q))$ is realistic when the transition model is sufficiently smooth; see [66] for an example of MDP satisfying this condition.

Assumption 6. *Let $\mathcal{Q}_M = \{Q \in \mathcal{Q} : J_1(Q) \le M\}$ and $\mathcal{G}_M = \{g \in \mathcal{G} : J_2(g) \le M\}$. There exists some constants C and $\alpha \in (0,1)$ such that for any $\epsilon, M > 0$,*

$$\max\left(\log \mathcal{N}_\infty(\epsilon, \mathcal{G}_M), \log \mathcal{N}_\infty(\epsilon, \mathcal{Q}_M)\right) \le C\left(\frac{M}{\epsilon}\right)^{2\alpha}$$

The sup-norm entropy conditions are satisfied for most common function class, e.g., Sobolev space and various RKHS; see [71, 72, 73, 74]. Here we use a common $\alpha \in (0,1)$ for both $\mathcal{Q}$ and $\mathcal{G}$ to simply the proof.

Theorem 1 (Global Convergence Rate)**.** *Suppose Assumption 1-6 hold. Let $(\hat{\eta}_n, \hat{Q}_n)$ be the estimator defined in (5.7). Then, for the tuning parameters (λ_n, μ_n) satisfying (i) $\mu_n = O_P(\lambda_n)$ and (ii) $\mu_n^{-1} = O_P(n^{-1/(1+\alpha)})$, we have*

$$\left\|\mathcal{E}(\cdot,\cdot;\hat{\eta}_n, \hat{Q}_n)\right\|_{\bar{\nu}_T}^2 = O_P(\lambda_n),\ J_1(\hat{Q}_n) = O_P(1).$$

Remark. In Lemma 5, we will show that up to a constant $|\hat{\eta}_n - \eta^\pi| \lesssim \|\mathcal{E}(\cdot,\cdot;\hat{\eta}_n,\hat{Q}_n)\|^2_{\bar{\nu}_T}$ and thus when $\lambda_n = o_P(1)$ we see that $\hat{\eta}_n$ is consistent estimate of η^π. On the other hand, we show that $\|(\mathcal{I}-\mathcal{P}^\pi)(\hat{Q}_n - Q^\pi)\|_{\bar{\nu}_T} \lesssim \|\mathcal{E}(\cdot,\cdot;\hat{\eta}_n,\hat{Q}_n)\|_{\bar{\nu}_T}$ where $\mathcal{P}^\pi$ is the conditional expectation operator $\mathcal{P}^\pi Q(s,a) = \mathbb{E}[\sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1},a')|S_t = a, A_t = a]$ and $\mathcal{I}$ is the identity operator. Although the main focus of this work is on the inference of average reward parameter, under suitable contraction assumptions on the transition kernel (see [75] for a similar analysis in discounted reward setting), one can relate this to the estimation error of the value function $\|\hat{Q}_n - \tilde{Q}^\pi\|_{\bar{\nu}_T}$ (recall that we can only estimate the value function up to a universal constant and thus $\tilde{Q}^\pi$ is used here). When the tuning parameters are chosen optimally, i.e., $\lambda_n \asymp \mu_n$ and $\lambda_n \asymp n^{-1/(1+\alpha)}$, we have $\|\mathcal{E}(\cdot,\cdot;\hat{\eta}_n,\hat{Q}_n)\|^2_{\bar{\nu}_T} = O_P(n^{-1/(1+\alpha)})$.

See the proof in Appendix 5.7. In what follows, we focus on the asymptotic of the estimator of average reward. In particular, we show that under certain assumptions the estimates $\hat{\eta}_n$ from (5.7) is $\sqrt{n}$ consistent when the tuning parameters are chosen appropriately and derive the asymptotic distribution. Recall $d^\pi(s)$ is the density of stationary distribution of the states under the target policy π and $\bar{d}_T(s,a)$ is the density of $\bar{\nu}_T$, the distribution of the states-action pair averaged across T decision times. We introduce some additional notations. Let $d^\pi(s,a) = \pi(a|s)d^\pi(s)$ be the stationary distribution of state-action under policy π. We define $e^\pi(s,a)$ by

$$e^\pi(s,a) = \frac{d^\pi(s,a)/\bar{d}_T(s,a)}{\int (d^\pi(s,a)/\bar{d}_T(s,a))d^\pi(s,a)dsda} \tag{5.9}$$

Note that e^π is a scaled version of an importance weight $d^\pi(s,a)/\bar{d}_T(s,a)$. The denominator is the expectation of the important weight under the stationary distribution and it is greater than 1:

$$\begin{aligned}
\int \frac{d^\pi(s,a)}{\bar{d}_T(s,a)} d^\pi(s,a)dsda &= \int \left[\frac{d^\pi(s,a)}{\bar{d}_T(s,a)}\right]^2 \bar{d}_T(s,a)dsda \\
&= \mathbb{E}\big[(1/T)\sum_{t=1}^T (d^\pi(S_t,A_t)/\bar{d}_T(S_t,A_t))^2\big] \\
&= \mathrm{Var}\big[(1/T)\sum_{t=1}^T (d^\pi(S_t,A_t)/\bar{d}_T(S_t,A_t))^2\big] + 1
\end{aligned}$$

where in the last equality the variance is with respect to the distribution of state-action under the behavioral policy in the training set and we use the fact that $\mathbb{E}[(1/T)\sum_{t=1}^T d^\pi(S_t,A_t)/\bar{d}_T(S_t,A_t)] = 1$ The definition of e^π is motivated by the least favorable direction in regression problems [71, 72].

Specifically, the use of importance weight, $d^\pi(s,a)/\bar{d}_T(s,a)$ in e^π implies that for all function Q

$$\mathbb{E}\Big[(1/T)\sum_{t=1}^{T}\big(Q(S_t,A_t)-\sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1},a')\big)e^\pi(S_t,A_t)\Big]=0 \tag{5.10}$$

Next, we introduce $q^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=1}^{\infty}(1-e^\pi(S_t,A_t))\,|\,S_1=s,A_1=a]$. Note that q^π has a similar structure to the value function in (5.3) and is well-defined as the long term average of $(1-e^\pi)$ is zero due to the denominator in e^π, i.e., $\int e^\pi(s,a)d^\pi(s,a)dsda = 1$. The construction of q^π allows us to rewrite $1-e^\pi(s,a)$ as the following:

$$1-e^\pi(s,a)=q^\pi(s,a)-\mathbb{E}\Big[\sum_{a'}\pi(S_{t+1},a')q^\pi(S_{t+1},a')\,|\,S_t=s,A_t=a\Big] \tag{5.11}$$

Assumption 7. *The shifted function $\tilde{q}^\pi \in \mathcal{Q}$ and $e^\pi \in \mathcal{G}$, where $\tilde{q}^\pi = q^\pi - q^\pi(s^*,a^*)$.*

Similar to the partially linear regression setting, the condition $\tilde{q}^\pi \in \mathcal{Q}$ is imposed to control the bias of $\hat{\eta}_n$, caused by penalization on the nonparametric component, e.g., $\hat{Q}_n$. In addition, we also require e^π is smooth enough (e.g., $e^\pi \in \mathcal{G}$) to control the bias due to not knowing the conditional expectation $\mathbb{E}_\pi[q^\pi(S_{t+1},A_{t+1})\,|\,S_t=s,A_t=a]$.

Theorem 2 (Asymptotic Normality)**.** *Suppose the conditions in Theorem 1 hold. In addition, suppose Assumptions 7 holds and $\lambda_n = o_P(n^{-1/2})$, the estimates $\hat{\eta}_n$ defined in (5.8) is $\sqrt{n}$-consistency and asymptotic Normal: $\sqrt{n}(\hat{\eta}_n-\eta^\pi)\Rightarrow N(0,\sigma^2)$, where*

$$\sigma^2=\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{d^\pi(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\big(R_{t+1}+\sum_{a'}Q^\pi(S_{t+1},a')-\eta^\pi-Q^\pi(S_t,A_t)\big)\right)$$

From Theorem 2, we see the variance in estimating the average reward parameter η^π depends on the importance weight between the stationary distribution of state-action pair induced by the target policy and the average state-action distribution in the training data. The closer of these two distributions implies a smaller variance of estimating the average reward. In the special case where the target policy is same as the behavior policy (i.e., on-policy evaluation) and the states in the training data follows the stationary distribution (e.g., when the length of trajectory is sufficiently large), one should expect to see the smallest variance. Although here we only focus on the asymptotic property of $\hat{\eta}_n$ for large n (n is the number of i.i.d. trajectories), one can see that increasing length of trajectory, T reduces the variance, as inside of the variance is an average over T decision time points.

To construct the confidence interval of $\hat{\eta}_n$, we need to the asymptotic variance σ^2. This can be done by plugging in $(\hat{\eta}_n, \hat{Q}_n)$ and an estimate of the importance weight For simplicity, denote the importance weight by $w^\pi(s,a) = d^\pi(s,a)/\bar{d}_T(s,a)$. Taking expectation on both side of (5.9) implies $w^\pi(s,a) = e^\pi(s,a)/\mathbb{E}[(1/T)\sum_{t=1}^T e^\pi(S_t,A_t)]$. Denote the estimates of e^π by $\hat{e}_n$. We can then estimate $w^\pi(s,a)$ by $\hat{e}_n(s,a)/\mathbb{P}_n[(1/T)\sum_{t=1}^T \hat{e}_n(S_t,A_t)]$. Motivated by the orthogonality (5.10), we estimate e^π by $\hat{g}_n(\hat{q}_n)$, where

$$\hat{q}_n = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathbb{P}_n[(1/T)\sum_{t=1}^T \hat{g}_n^2(S_t,A_t;q)] + \lambda_n J_1^2(q)$$

$$q\in\mathcal{Q},\ \hat{g}_n(\cdot,\cdot;q) = \operatorname*{argmin}_{g\in\mathcal{G}} \mathbb{P}_n[(1/T)\sum_{t=1}^T \big(U(S_t,A_t,S_{t+1};q) - g(S_t,A_t)\big)^2] + \mu_n J_2^2(g)$$

and $U(s,a,s';q) = 1 - q(s,a) + \sum_{a'}\pi(a'|a)q(s',a')$. It is easy to verify that this is a consistent estimates of σ^2 under the assumptions listed in Theorem 2. In Appendix 5.7, we provide a closed-form solution of the estimates asymptotic variance when $\mathcal{Q}$ and $\mathcal{G}$ are RKHSs.

5.5 Extensions

In this section, we first develop two generalizations of previous results. We first generalize the policy evaluation of a single policy to a class of policies and derive the asymptotic of the estimated average rewards. We then consider the setting where there is a time-invariant state and the average reward of the target policy depends on the time-variant state. We end with a discussion on how to build the estimates based on the state-only function when the behavior policy is known.

Policy evaluation of multiple policies

Let Π be a class of target policies. The policy class can be a finite collection of policies of interest or a class of policy parameterized by some parameters. Denote by $\hat{\eta}_n^\pi$ the estimated average reward of policy $\pi\in\Pi$ from (5.7). Assume the entropy integral is finite: $\int_0^\infty (\log \mathcal{N}_\infty(\epsilon,\Pi)^{1/2} d\epsilon < \infty$. In addition, suppose for all policy $\pi\in\Pi$, the Assumption 1-8 are satisfied and $\sup_{\pi\in\Pi} J_1(Q^\pi) < \infty$. The proof of Theorem 1 and 2 can be easily extended to hold simultaneously for all policy $\pi\in\Pi$ using the finite entropy integral and we can show that the estimated average reward converges in distribution to a Gaussian process indexed by the policy. In particular, we have $\sqrt{n}(\hat{\eta}_n^\pi - \eta^\pi) \Rightarrow \mathbb{G}(\pi)$ in $l^\infty(\Pi)$, where $\mathbb{G}$ is mean zero Gaussian process indexed by π with covariance function $\mathbb{E}[(1/T^2)\sum_{t=1}^T \delta_t^{\pi_1} w_t^{\pi_1} \delta_t^{\pi_2} w_t^{\pi_2}]$ for $\pi_1,\pi_2\in\Pi$. Here $\delta_t^\pi = R_{t+1} + \sum_{a'}\pi(a'|S_{t+1})Q^\pi(S_{t+1},a') -$

$\eta^\pi - Q^\pi(S_t.A_t)$ and $w_t^\pi = w^\pi(S_t, A_t) = d^\pi(S_t, A_t)/\bar{d}_T(S_t, A_t)$.

This result can be used to compare two policies. For example, we can form a confidence interval of $(\eta^{\pi_1} - \eta^{\pi_2})$, the difference of the average reward. On the other hand, consider a policy class parameterized by β in a compact set $\mathcal{B}$ for binary action, e.g., $\pi_\beta(1|s) = \exp(\beta^\top f(s))/(1 + \exp(\beta^\top f(s)))$. Using the result above, we can form a $(1-\alpha)\%$ simultaneous confidence band $\{\hat{\eta}_n^{\pi_\beta} \pm n^{-1/2} z_\alpha : \beta \in \mathcal{B}\}$ for the average reward where the quantile z_α is determined such that $\Pr(\sup_{\beta\in\mathcal{B}} |\mathbb{G}(\pi_\beta)| \le z_\alpha) = 1 - \alpha$

Incorporating time-invariant states

In mobile health application, often the baseline demographic information, e.g., gender and occupation are collected. However including time-invariant information into the states would violate Assumption 1. In the followings, we extend the previous result by incorporating the baseline information. Let $S_t = (Z, X_t)$ where $Z \in \mathcal{Z}$ is the time-invariant baseline information and $X_t \in \mathcal{X}$ is the time-varying variables. We consider a target policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. Note that the input of the target policy can depend on the baseline.

We generalize Assumption 1 with respect to the target policy π as follows. We assume that for all $z \in \mathcal{Z}$, the induced markov chain on $\mathcal{X}$ by target policy π with the transition density $p(x'|x, a) = \sum_a \pi(a|x, z)p(x'|x, a, z)$ is irreducible and aperiodic for all $z \in \mathcal{Z}$. The density of stationary distribution exists for each $z \in \mathcal{Z}$ is denoted by $d^\pi(x|z)$. Under this assumption, the average reward defined in (5.1) can be shown to be a function of baseline variable z only, i.e., $\eta^\pi(s) = \eta^\pi(z)$ and the value function can be identified up to a function of z by solving Bellman equation:

$$\mathbb{E}_\pi[R_{t+1} + Q(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a] = \eta(z) + Q(s, a). \tag{5.12}$$

Furthermore, we assume that the average reward, as a function of z, follows a linear model, i.e., $\eta^\pi(z) = f(z)^\top \beta^\pi$ where $f(z)$ is a feature vector of length p. Similar to the estimator presented in Section 5.3, we can estimate β^π, as well as the shifted value function $\tilde{Q}^\pi(s, a) = Q^\pi(s, a) - Q^\pi((z, x^*), a^*)$ for some reference time-varying state x^* and action a^*, by

$$(\hat{\beta}_n, \hat{Q}_n) = \underset{(\beta,Q)\in\mathbb{R}^p\times\mathcal{Q}}{\operatorname{argmin}} \mathbb{P}_n \left[\frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{E}}_n^2(S_t, A_t; \beta, Q)\right] + \lambda_n J_1^2(Q) \tag{5.13}$$

where the state-action function $\hat{\mathcal{E}}_n(\cdot,\cdot\beta,Q)$ is given by

$$\hat{\mathcal{E}}_n(\cdot,\cdot;\beta,Q) = \underset{g\in\mathcal{G}}{\operatorname{argmin}}\,\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\big(\delta_t(\beta,Q)-g(S_t,A_t)\big)^2\right]+\mu_n J_2^2(g) \tag{5.14}$$

where $\delta_t(\beta,Q) = R_{t+1}+\sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1},a')-f(Z)^\top\beta-Q(S_t,A_t)$ is the temporal difference error. The global rate of convergence of the estimator $(\hat{\beta}_n,\hat{Q}_n)$ can be shown similarly as in the proof of Theorem 1 and thus skipped here. To derive the asymptotic of $\hat{\beta}_n$, the definition of e^π in (5.9) can be generalized as follow. For $s=(z,x)$, define the vector

$$e^\pi(s,a) = (e_1^\pi(s,a),\ldots,e_p^\pi(s,a))^\top = \frac{d^\pi(x,a|z)/\bar{d}_T(x,a|z)}{\int(d^\pi(x,a|z)/\bar{d}_T(x,a|z))d^\pi(x,a|z)dxda}f(z)\in\mathbb{R}^p$$

where $d^\pi(x,a|z)=\pi(a|x,z)d^\pi(x|z)$ is the density of stationary distribution of the time-varying states and action given the baseline z with respect to the target policy π and similarly $\bar{d}_T(x,a|z)$ is the density of the average distribution of time-varying states and action of the trajectory $\mathcal{D}$ given the baseline z. Under the similar set of conditions in Theorem 2, it can be shown that

$$\sqrt{n}(\hat{\beta}_n-\beta^\pi)\Rightarrow N(0,U^{-1}VU^{-\top})$$

where $U=\mathbb{E}[\sum_{t=1}^T e^\pi(S_t,A_t)f(Z)^\top]$ and $V=\mathrm{Var}[\sum_{t=1}^T\delta_t^\pi e^\pi(S_t,A_t)]$ with the temporal difference error $\delta_t^\pi = R_{t+1}+\sum_{a'}\pi(a'|S_{t+1})Q^\pi(S_{t+1},a')-f(Z)^\top\beta^\pi-Q^\pi(S_t,A_t)$.

Similar to the estimate of asymptotic variance in Section 5.4, we can form a sandwich estimator, $\hat{U}^{-1}\hat{V}\hat{U}^{-\top}$. Firstly, estimate e^π by $\hat{e}_n=(\hat{g}_{n,1}(\hat{q}_{n,1}),\ldots,\hat{g}_{n,p}(\hat{q}_{n,p}))^\top$, where $\hat{g}_{n,k}(\cdot,\cdot;q)=\operatorname{argmin}_{g\in\mathcal{G}}\mathbb{P}_n[\frac{1}{T}\sum_{t=1}^T(f_k(Z)-q(S_t,A_t)+\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')-g(S_t,A_t))^2]+\mu_n J_2^2(g)$ and $\hat{q}_{n,k}=\operatorname{argmin}_{q\in\mathcal{Q}}\mathbb{P}_n[(1/T)\sum_{t=1}^T\hat{g}_{n,k}^2(S_t,A_t;q)]+\lambda_n J_1^2(q)$. Then we estimate U and V by

$$\hat{U}=\mathbb{P}_n[\sum_{t=1}^T\hat{e}_n(S_t,A_t)f(Z)^\top],\ \hat{V}=\mathbb{P}_n[\big(\sum_{t=1}^T\hat{\delta}_t\hat{e}_n(S_t,A_t)\big)\big(\sum_{t=1}^T\hat{\delta}_t\hat{e}_n(S_t,A_t)\big)^\top]$$

where $\hat{\delta}_t$ is the plug-in estimates. In Appendix 5.7, we provide closed-form formula of $(\hat{\beta}_n,\hat{Q}_n,\hat{e}_n)$ when using RKHS.

Estimating state value function

So far we've only considered estimating the average reward η^π through the state-action value function $Q^\pi(s,a)$. And note that the method does not require access to the behavior policy that chooses

the actions in the training set. In the case where the behavior policy is known (for example the data is collected from Micro-randomized trial), one can build a similar estimator based on state-only value function $V^\pi(s) = \sum_a \pi(a|s)Q^\pi(s,a)$ by using the importance weight.

In this case, similar to Assumption 1, Bellman Equation for the state value function becomes $\mathbb{E}_\pi[R_{t+1} + V(S_{t+1}) \,|\, S_t = s] = \eta + V(s)$ for all s. Note that the expectation in the RHS is taken under $A_t \sim \pi(\cdot|S_t)$. Consider two classes of function of state $\mathcal{V}$ and $\mathcal{G}$ with regularizer J_1, J_2. Similar to the estimator presented in (5.7), we can estimate $(\eta^\pi, \tilde{V}^\pi)$ by

$$(\hat{\eta}_n, \hat{V}_n) = \underset{(\eta,V)\in\mathbb{R}\times\mathcal{V}}{\operatorname{argmin}} \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{E}}_n^2(S_t;\eta,V)\big] + \lambda_n J_1^2(V)$$

$$\forall(\eta,V),\ \hat{\mathcal{E}}_n(\cdot;\eta,V) = \underset{g\in\mathcal{G}}{\operatorname{argmin}}\, \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \frac{\pi(A_t|S_t)}{\pi_b(A_t|H_t)}\big(\delta_t(\eta,V) - g(S_t)\big)^2\big] + \mu_n J_2^2(g)$$

where $\delta_t(\eta,V) = R_{t+1} + V(S_{t+1}) - \eta - V(S_{t+1})$. Theorem 1 and 2 can be extended to this estimator under similar assumptions. In particular, $\sqrt{n}(\hat{\eta}_n - \eta^\pi) \Rightarrow N(0,\sigma^2)$ and σ^2 is given by

$$\sigma^2 = \text{Var}\Big[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\pi(A_t|S_t)}{\pi_b(A_t|H_t)}\right)\left(\frac{d^\pi(S_t)}{\bar{d}_T(S_t)}\right)\big(R_{t+1} + V^\pi(S_{t+1}) - \eta^\pi - V^\pi(S_t)\big)\Big]$$

where d^π and $\bar{d}_T$ is the density of the stationary distribution and average state distribution in the trajectory $\mathcal{D}$, respectively. The advantage of this approach is that we only need to estimate the state value function, which is an average of the state-action value function and thus one can expect it reducing variance. When behavior policy is very different from target policy, the variance might become larger. Further investigation of this trade-off is left for future work.

5.6 Discussion

In this work, we developed a flexible method to estimate the long-term average outcome for a given treatment policy. The method uses a non-parametric function class to model the value function and we developed the asymptotic property of the estimated average reward. In mobile health application, one big challenge is the non-stationarity due to unobserved aspects of the current context (e.g., the engagement and/or burden). For example, the mapping from states to reward is likely to be different over time. It will be interesting to generalize the current framework of long-term average reward to the non-stationary setting. Alternatively, one can consider evaluating the treatment policy in the indefinite horizon setting where there is an absorbing state (akin to

the user stop using the mobile app) and consider estimating the total rewards till the absorbing state is reached. While evaluating a given treatment policy or a class of treatment policies1 is an important first step towards developing efficient JITAIs, it is crucial to develop methods for estimating the optimal policy, which leads to the largest positive health behavior outcomes, as well as the inferential method for assessing the usefulness of certain states in the optimal treatment policy.

5.7 Appendix

A. Computation of Estimator and Asymptotic Variance

Below, we derive the closed-form solution of the estimators as well as the asymptotic variance when $\mathcal{Q}$ and $\mathcal{G}$ are both Reproducing Kernel Hilbert Space (RKHS). We consider the general setting discussed in Section 5.5 where there is time-invariant information in the state and the average reward is modeled as $f(z)^\top \beta$ for some feature vector f. The case where there is no time-invariant state can be covered by setting $f = 1$. Without loss of generality, we denote the training data simply by $\mathcal{D} = \{Z_h, X_h, A_h, X'_h, R_i\}_{i=1}^N$ where h indexes the tuple of transition sample with baseline in the training set, Z_h is the corresponding baseline, X_h and X'_h is the current and next time-varying state and R_h is the reward. Let $U_h = (Z_h, X_h, A_h)$ be the state-action pair, $S'_h = (Z_h, X'_h)$ and $U'_h = (Z_h, X_h, A_h, X'_h)$.

It is more common to form the RKHS for the function of state. Suppose the kernel function for the state function is given by $k_0(s_1, s_2)$, $s_1, s_2 \in \mathcal{S}$. To incorporate the action (assume binary below), one can define $k((s_1, a_1), (s_2, a_2)) = \mathbb{1}_{\{a_1=a_2\}} k_(s_1, s_2)$. That is, we model each arm separately with each $Q(\cdot, a)$ in the RKHS with kernel k_0. Alternatively, one can also model the baseline value $Q(s, 0)$ and the difference, i.e., $Q(s, 1) - Q(s, 0)$ with two kernels k_0, k_1. In this case, one can define the kernel by $k((s_1, a_1), (s_2, a_2)) = k_0(s_1, s_2) + \mathbb{1}_{\{a_1=a_2=1\}} k_1(s_1, s_2)$.

Suppose the kernel function for $\mathcal{Q}$ and $\mathcal{G}$ are given by $k(\cdot, \cdot), l(\cdot, \cdot)$. Denote the inner product by $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$, respectively. Recall that we need to restrict the function space $\mathcal{Q}$ such that $Q((z, x^*), a^*) = 0$ for all $Q \in \mathcal{Q}$. For an arbitrary kernel function k_0 on $\mathcal{S} \times \mathcal{A}$, we can always transform it into k such that this assumption is satisfied. In particular, define $k(U_1, U_2) = k_0(U_1, U_2) - k_0((Z_1, x^*, a^*), U_2) - k_0(U_1, (Z_2, x^*, a^*)) + k_0((Z_1, x^*, a^*), (Z_2, x^*, a^*))$. It can be seen that the induced RKHS by k satisfies the condition.

Note that the minimizer (5.14) is a regression problem. And it is well known that the solution is given by $\hat{\mathcal{E}}_n(\eta, Q) = \sum_{h=1}^N l(U_i, \dot{)} \theta_h(\eta, Q)$ where $\theta(\eta, Q) = (L + \mu I_N)^{-1} \delta_N(\eta, Q)$, where

$l_U = l(U, \cdot)$, $\mu = \mu_n N$ and $\delta_N(\eta, Q) = (\delta(U'_h; \eta, Q))$ is the vector of TD error. In addition, each TD error can be written as $\delta(U'; \beta, Q) = R - f(Z)^\top \beta - \langle Q, \tilde{k}_{U'} \rangle_{\mathcal{G}}$ where

$$\tilde{k}_{U'}(\cdot) = k(U, \cdot) - \sum_{a'} \pi(a'|S')k((S', a'), \cdot) \in \mathcal{Q}$$

We show that the solution of (5.13) must stay in the linear span: $\{\sum_{h=1}^N \alpha_h \tilde{k}_{U'_h}(\cdot) : \alpha_h \in \mathbb{R}, h = 1, \ldots, N\}$. To see this, suppose $Q = Q_0 + \Delta$ where $Q_0 = \sum_{h=1}^N \alpha_h \tilde{k}_{U'_h}$ and $\Delta \in \mathcal{Q}$ satisfies $\langle \Delta, \tilde{k}_{U'_h} \rangle_{\mathcal{Q}} = 0$ for all $h = 1, \ldots, N$. For the first term in (5.13), denoted by $L(\beta, Q)$, we have $L(\beta, Q) = L(\beta, Q_0)$, using the fact that each TD error is unchanged by adding Δ. And the second term, we have $\|Q\|^2_{\mathcal{Q}} = \|Q_0\|^2_{\mathcal{Q}} + \|\Delta\|^2_{\mathcal{Q}}$ due to the orthogonality of Δ. Thus the minimizer must have $\Delta = 0$. Using this finite representer property, we can find $(\hat{\beta}, \hat{\alpha})$ by

$$(\hat{\beta}, \hat{\alpha}) = \underset{\beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^N}{\operatorname{argmin}} (R_N - F\beta - \tilde{K}\alpha)^\top M (R_N - F\beta - \tilde{K}\alpha) + \lambda \alpha^\top \tilde{K} \alpha$$

where $R_N = (R_h)_{h=1}^N$, $\tilde{K} = (\langle \tilde{k}_{U'_h}, \tilde{k}_{U'_k} \rangle_{\mathcal{Q}})_{k,h=1}^N$, $M = (L + \mu I_N)^{-1} L^2 (L + \mu I_N)^{-1}$, $F = (f(Z_h))_{h=1}^N$ and $\lambda = \lambda_n N$. Note that we have $\tilde{K}[h, k]$ can be calculated by

$$\begin{aligned}\langle \tilde{k}_{U'_h}, \tilde{k}_{U'_k} \rangle_{\mathcal{Q}} = k(U_h, U_k) - \sum_{a'} \pi(a'|S'_h)k((S'_h, a'), U_k) - \sum_{a'} \pi(a'|S'_k)k((S'_k, a'), U_h) \\ + \sum_{a'_h} \sum_{a'_k} \pi(a'_h|S'_h)\pi(a'_k|S'_k)k((S'_h, a'_h), (S'_k, a'_k))\end{aligned}$$

Taking derivative implies $(\hat{\beta}, \hat{\alpha})$ solves

$$\begin{aligned}F^\top M F \beta &= F^\top M (R_N - \tilde{K}\alpha) \\ (M\tilde{K} + \lambda I_N)\alpha &= M(R_n - F\beta)\end{aligned}$$

Similarly, we can find the closed-form solution of $\hat{e} = (\hat{e}_k)_{k=1}^p$. For each k, $\hat{e}_k = \sum_{h=1}^N \theta_h l(U_h, \cdot)$ where $\theta = (\theta_h)_{h=1}^N = (L + \mu I_N)^{-1}(F_k - \tilde{K}\hat{\alpha})$ and $\hat{\alpha}$ solves $(M\tilde{K} + \lambda I_N)\alpha = MF_k$. Here F_k is the k-the column of F.

B. Proof of Theorem 1 and 2

In the following, for simplicity we use $\mathcal{E}(\eta, Q)$ to denote the state-action function $\mathcal{E}(\cdot, \cdot; \eta, Q)$ and similar for $\mathcal{E}_{\mathcal{G}}(\eta, Q)$, $\hat{\mathcal{E}}_n(\eta, Q)$.

Proof of Theorem 1. By the definition of κ in Assumption 4,

$$\|\mathcal{E}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 \leq \frac{1}{\kappa^2}\|\mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 \leq \frac{2}{\kappa^2}\left(\|\mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n) - \hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 + \|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2\right)$$

The second term is bounded by Lemma 4. We focus on the first term. From Lemma 3, with probability $1-\delta$, for some constant c_1 the first term

$$\begin{aligned}&\|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n) - \mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 \\ &\leq c_1(1 + J_1^2(\hat{Q}_n) + J_2^2(\mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n)) + \log(1/\delta))\mu_n\end{aligned}$$

Using Assumption 5 with the constants C_1, C_2, we have

$$\|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n) - \mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 \leq c_1(1 + 2C_1^2 + (1 + 2C_2^2)J_1^2(\hat{Q}_n) + \log(1/\delta))\mu_n \tag{5.15}$$

To bound $J_1^2(\hat{Q}_n)$, the optimizing property of the estimators $(\hat{\eta}_n, \hat{Q}_n)$ in (5.7) and Lemma 3 imply that

$$\begin{aligned}\lambda_n J_1^2(\hat{Q}_n) &\leq \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{E}}_n^2(S_t, A_t; \hat{\eta}_n, \hat{Q}_n)\right] + \lambda_n J_1^2(\hat{Q}_n) \\ &\leq \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{E}}_n^2(S_t, A_t; \eta^\pi, \tilde{Q}^\pi)\right] + \lambda_n J_1^2(\tilde{Q}^\pi) \\ &\leq \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(\hat{\mathcal{E}}_n(S_t, A_t; \eta^\pi, \tilde{Q}^\pi) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta^\pi, \tilde{Q}^\pi))^2\right] + \lambda_n J_1^2(\tilde{Q}^\pi) \\ &\leq c_1(1 + J_1^2(\tilde{Q}^\pi) + \log(1/\delta))\mu_n + \lambda_n J_1^2(\tilde{Q}^\pi)\end{aligned}$$

Since $\mu_n = O_P(\lambda_n)$, we have $J_1^2(\hat{Q}_n) \leq c_2(1 + J_1^2(\tilde{Q}^\pi) + \log(1/\delta))$ for some constant c_2. Combining with (5.15) gives $\|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n) - \mathcal{E}_{\mathcal{G}}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}^2 \leq c(\delta)\mu_n$ for some $c(\delta) > 0$.

□

Lemma 3. *Let $\mathcal{E}_{\mathcal{G}}(\eta, Q)$ be the surroage Bellman error operator defined in (5.6). Let $\hat{\mathcal{E}}_n(\eta, Q)$ be the estimated surrogate Bellman error in (5.5) with tuning parameter μ_n. Suppose Assumptions 2, 3,4 and 6 hold. Then with probability at least $1-\delta$, the followings hold up to some constant for all $\eta \in [-R_{\max}, R_{\max}]$ and $Q \in \mathcal{Q}$:*

$$\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_{\bar{\nu}_T}^2 \lesssim \mu_n(J_1^2(Q) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q))) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}$$

$$J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) \lesssim J_1^2(Q) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + \frac{1}{n\mu_n^{(\alpha+1)}} + \frac{1}{n\mu_n} + \frac{\log(1/\delta)}{n\mu_n}$$
$$\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_n^2 \lesssim \mu_n J_1^2(Q) + \mu_n J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + \frac{1}{n\mu_n^{\alpha}} + \frac{1}{n} + \frac{\log(1/\delta)}{n}$$

where $\|g\|_n^2 = \mathbb{P}_n[1/T \sum_{t=1}^T g^2(S_t, A_t)]$ *is the empirical 2-norm. As a result when* $\mu_n^{-1} = O_P(n^{\frac{1}{1+\alpha}})$*, with probability at least* $1-\delta$*, for all* (η, Q)

$$\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_{\bar{\nu}_T}^2 \lesssim (1 + J_1^2(Q) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + \log(1/\delta))\mu_n$$
$$J_2(\hat{\mathcal{E}}_n(\eta, Q)) \lesssim J_1(Q) + J_2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + \sqrt{\log(1/\delta)}$$
$$\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_n^2 \lesssim (1 + J_1^2(Q) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + \log(1/\delta))\mu_n$$

Proof of Lemma 3. For simplicity, let $\delta_t(\eta, Q) = \delta(S_t, A_t, R_{t+1}, S_{t+1}; \eta, Q)$ be the temporal difference error at time t with respect to (η, Q). We start with decomposing the error $\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_{\bar{\nu}_T}^2$:

$$\begin{aligned}
\|\hat{\mathcal{E}}_n(\eta, Q) - \mathcal{E}_{\mathcal{G}}(\eta, Q)\|_{\bar{\nu}_T}^2 &= (1/T) \sum_{t=1}^T \mathbb{E}[(\hat{\mathcal{E}}_n(S_t, A_t; \eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))^2] \\
&= (1/T) \sum_{t=1}^T \mathbb{E}[(\hat{\mathcal{E}}_n(S_t, A_t; \eta, Q) - \delta_t(\eta, Q) + \delta_t(\eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))^2] \\
&= (1/T) \sum_{t=1}^T \mathbb{E}[(\delta_t(\eta, Q) - \hat{\mathcal{E}}_n(S_t, A_t; \eta, Q))^2] + (1/T) \sum_{t=1}^T \mathbb{E}[(\delta_t(\eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))^2] \\
&\quad + (2/T) \sum_{t=1}^T \mathbb{E}[(\hat{\mathcal{E}}_n(S_t; \eta, Q) - \delta_t(\eta, Q))(\delta_t(\eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))]
\end{aligned}$$

Since $\sum_{t=1}^T \mathbb{E}[(\mathcal{E}(S_t, A_t; \eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))g(S_t, A_t)] = 0$ for all $g \in \mathcal{G}$ due to the optimizing property of $\mathcal{E}_{\mathcal{G}}$ in (5.6), we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}[(\hat{\mathcal{E}}_n(S_t; \eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q) + \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q) - \delta_t(\eta, Q))(\delta_t(\eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))] \\
&= \sum_{t=1}^T \mathbb{E}[(\mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q) - \delta_t(\eta, Q))(\delta_t(\eta, Q) - \mathcal{E}_{\mathcal{G}}(S_t, A_t; \eta, Q))]
\end{aligned}$$

$$= -2\sum_{t=1}^{T}\mathbb{E}[(\delta_t(\eta,Q)-\mathcal{E}_{\mathcal{G}}(S_t,A_t;\eta,Q))^2]$$

Thus we have

$$\|\hat{\mathcal{E}}_n(\eta,Q)-\mathcal{E}_{\mathcal{G}}(\eta,Q)\|^2_{\tilde{\nu}_T}=\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\delta_t(\eta,Q)-\hat{\mathcal{E}}_n(S_t;\eta,Q))^2-(\delta_t(\eta,Q)-\mathcal{E}_{\mathcal{G}}(S_t,A_t;\eta,Q))^2\right]$$

For $g_1,g_2\in\mathcal{G},\eta\in\mathbb{R},Q\in\mathcal{Q}$, define the function of trajectory $\mathcal{D}=\{S_1,A_1,\ldots,S_{T+1}\}$

$$f(g_1,g_2,\eta,Q):\mathcal{D}\mapsto\frac{1}{T}\sum_{t=1}^{T}(\delta_t(\eta,Q)-g_1(S_t,A_t))^2-(\delta_t(\eta,Q)-g_2(S_t,A_t))^2$$

and $J^2(g_1,g_2,V)=J_2^2(g_1)+(2/3)J_2^2(g_2)+(2/3)J_1^2(Q)$. Let $\|g\|_n^2=\mathbb{P}_n[1/T\sum_{t=1}^{T}g^2(S_t,A_t)]$. With these notations, we have

$$\begin{aligned}
&\|\hat{\mathcal{E}}_n(\eta,Q)-\mathcal{E}_{\mathcal{G}}(\eta,Q)\|^2_{\tilde{\nu}_T}+\|\hat{\mathcal{E}}_n(\eta,Q)-\mathcal{E}_{\mathcal{G}}(\eta,Q)\|_n^2+\mu_nJ_2^2(\hat{\mathcal{E}}_n(\eta,Q))\\
&=Pf(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),\eta,Q)+\mathbb{P}_nf(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),\eta,Q)+\mu_nJ_2^2(\hat{\mathcal{E}}_n(\eta,Q))\\
&=I_1(\eta,Q)+I_2(\eta,Q)
\end{aligned}$$

where $I_1(\eta,Q)=3(\mathbb{P}_nf(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),\eta,Q)+\mu_nJ^2(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),V))$ and $I_2(\eta,Q)=(\mathbb{P}_n+P)f(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),\eta,Q)+\mu_nJ_2^2(\hat{\mathcal{E}}_n(\eta,Q))-I_1(\eta,Q)$.

For the first term, the optimizing property of $\hat{\mathcal{E}}_n(\eta,Q)$ implies that

$$\begin{aligned}
&(1/3)I_1(\eta,Q)=\mathbb{P}_nf(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),\eta,Q)+\mu_nJ^2(\hat{\mathcal{E}}_n(\eta,Q),\mathcal{E}_{\mathcal{G}}(\eta,Q),V)\\
&=\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\delta_t(\eta,Q)-\hat{\mathcal{E}}_n(S_t;\eta,Q))^2-(\delta_t(\eta,Q)-\mathcal{E}_{\mathcal{G}}(S_t,A_t;\eta,Q))^2\big]\\
&\quad+\mu_nJ_2^2(\hat{\mathcal{E}}_n(\eta,Q))+(2/3)\mu_nJ_2^2(\mathcal{E}_{\mathcal{G}}(\eta,Q))+(2/3)\mu_nJ_1^2(Q)\\
&=[\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\delta_t(\eta,Q)-\hat{\mathcal{E}}_n(S_t;\eta,Q))^2\big]+\mu_nJ^2(\hat{\mathcal{E}}_n(\eta,Q))]\\
&\quad-\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\delta_t(\eta,Q)-\mathcal{E}_{\mathcal{G}}(S_t,A_t;\eta,Q))^2\big]+(2/3)\mu_nJ_2^2(\mathcal{E}_{\mathcal{G}}(\eta,Q))+(2/3)\mu_nJ_1^2(Q)\\
&\leq(5/3)\mu_nJ_2^2(\mathcal{E}_{\mathcal{G}}(\eta,Q))+(2/3)\mu_nJ_1^2(Q)
\end{aligned}$$

Thus, $I_1(\eta,Q)\leq 5\mu_nJ_2^2(\mathcal{E}_{\mathcal{G}}(\eta,Q))+2\mu_nJ_1^2(Q)$ holds for all (η,Q).

Next, to obtain the uniform bound of $I_2(\eta, Q)$ over all η, Q, we apply the peeling device together with the exponential inequality of the relative deviation of the empirical process developed, Theorem 19.3 in [74]. This is similar to the proof of Lemma 15 in . Rewrite $I_2(\eta, Q)$ as

$$\begin{aligned}
I_2(\eta, Q) &= (\mathbb{P}_n + P) f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) + \mu_n J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) \\
&\quad - 3(\mathbb{P}_n f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) + \mu_n J^2(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), V)) \\
&= 2(P - \mathbb{P}_n) f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) - P f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) \\
&\quad - 2\mu_n (J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + J_1^2(Q))
\end{aligned}$$

Define the interval $B = [-R_{\max}, R_{\max}]$. Fix some $t > 0$.

$$\begin{aligned}
&\Pr\left(\exists (\eta, Q) \in B \times \mathcal{Q}, I_2(\eta, Q) > t\right) \\
&= \sum_{l=0}^{\infty} \Pr\Big(\exists (\eta, Q) \in B \times \mathcal{Q},\ 2\mu_n [J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + J_1^2(Q)] \in [2^l t \mathbb{1}_{\{l \neq 0\}}, 2^{l+1} t), \\
&\qquad 2(P - \mathbb{P}_n) f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) > P f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) \\
&\qquad + 2\mu_n [J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + J_1^2(Q)] + t\Big) \\
&\leq \sum_{l=0}^{\infty} \Pr\Big(\exists (\eta, Q) \in B \times \mathcal{Q},\ 2\mu_n [J_2^2(\hat{\mathcal{E}}_n(\eta, Q)) + J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) + J_1^2(Q)] \leq 2^{l+1} t, \\
&\qquad 2(P - \mathbb{P}_n) f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) > P f(\hat{\mathcal{E}}_n(\eta, Q), \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) + 2^l t\Big) \\
&\leq \sum_{l=0}^{\infty} \Pr\left(\sup_{f \in \mathcal{F}_l} \frac{(P - \mathbb{P}_n) f(\mathcal{D})}{P f(\mathcal{D}) + 2^l t} > \frac{1}{2}\right)
\end{aligned}$$

where $\mathcal{F}_l = \{f(g, \mathcal{E}_{\mathcal{G}}(\eta, Q), \eta, Q) : J_2^2(g) \leq \frac{2^l t}{\mu_n}, J_2^2(\mathcal{E}_{\mathcal{G}}(\eta, Q)) \leq \frac{2^l t}{\mu_n}, J_1^2(Q) \leq \frac{2^l t}{\mu_n}, \eta \in B\}$. Next we verify the conditions in Theorem 19.3 in [74] with $\mathcal{F} = \mathcal{F}_l$, $\epsilon = 1/2$ and $\eta = 2^{l+1} t$ to get an exponential inequality for each term in the summation, similar to the proof of Lemma 4 below. The (A1) and (A2) assumptions are easy to verify using Assumptions 2, 3 and 4. There exists some K_1, K_2 (depending on $\pi_{\min}$, $R_{\max}, Q_{\max}$) such that $\|f\|_\infty \leq K_1$ and $\mathbb{E}[f(\mathcal{D})^2] \leq K_2 \mathbb{E}[f(\mathcal{D})]$. To ensure the assumption A3 holds for every l, we just need to ensure $t > c_1 n^{-1}$ for some constant $c_1(K_1, K_2)$. Using the entropy condition, it can be shown that $\log \mathcal{N}(u, \mathcal{F}_l, \|\cdot\|_n) \leq c_2 \left(\frac{2^l t}{\mu_n}\right)^{\alpha} \log\left(\frac{4R_{\max} + u}{u}\right) u^{-2\alpha}$ for some constant c_2 depending on $R_{\max}, Q_{\max}$ and the constant in Assumption 6. Thus the condition A4 is satisfied for every l by choosing $t \geq c_3 (n \mu_n^{\alpha})^{-1}$. Therefore,

$$\Pr\left(\exists (\eta, Q) \in B \times \mathcal{Q}, I_2(\eta, Q) > t\right) \leq C_1 \exp(-C_2 n t)$$

By choosing $t \geq c_4[(n\mu_n^{\alpha})^{-1} + n^{-1} + \frac{\log(1/\delta)}{n}]$, we can see that

$$I_2(\eta, Q) \leq c_4[(n\mu_n^{\alpha})^{-1} + n^{-1} + \frac{\log(1/\delta)}{n}]$$

holds for all η, Q with probability at least $1 - \delta$.

□

Lemma 4. *Suppose the conditions in Lemma 3 and Assumptions 3, 5 hold. Let $(\hat{\eta}_n, \hat{Q}_n)$ be the estimator in (5.7) and $\hat{\mathcal{E}}_n(\eta, Q)$ be the estimated Bellman error operator in (5.8). With the tuning parameters such that $\mu_n = O_P(\lambda_n)$ and $\mu_n^{-1} = O_p(n^{1/(1+\alpha)})$, there exists some constant $c(\delta)$ such that the following holds with probability $1 - \delta$*

$$\|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)\|_{\tilde{\nu}_T}^2 \lesssim c(\delta)\lambda_n$$
$$\mathbb{P}_n\big[1/T \sum_{t=1}^{T} \hat{\mathcal{E}}_n^2(S_t, A_t; \hat{\eta}_n, \hat{Q}_n)\big] \leq c(\delta)\lambda_n$$

Proof of Lemma 4. Fix some $\delta > 0$. For ease of notation, define a function

$$f(g) : \mathcal{D} \to \frac{1}{T} \sum_{t=1}^{T} g(S_t, A_t)^2$$

for $g \in \mathcal{G}$. Let $\|g\|_n^2 = \mathbb{P}_n[1/T \sum_{t=1}^{T} g^2(S_t, A_t)]$. Then

$$\|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)\|_{\tilde{\nu}_T}^2 + \|\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)\|_n^2 = (P + \mathbb{P}_n) f(\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)) = I_{n,1} + I_{n,2}$$

where $I_{n,1} = 3(\mathbb{P}_n f(\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)) + (2/3)\lambda_n J_1^2(\hat{Q}_n))$ and $I_{n,2} = (\mathbb{P}_n + P) f(\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n)) - I_{n,1}$.

Without loss of generality, we assume the average reward estimates $\hat{\eta}_n \in [-R_{\max}, R_{\max}]$ (otherwise a truncation argument can be applied). For the first term, Assumption 3, the optimizing property (5.7) and the in-sample error bound in Lemma 3 with the constant C imply that under the choice of (μ_n, λ_n) specified in the condition, the following holds with probability at least $1 - \delta/2$ for some constant $c_0 > 0$

$$\begin{aligned} I_{n,1} &\leq 3\mathbb{P}_n f(\hat{\mathcal{E}}_n(\eta^{\pi}, \tilde{Q}^{\pi})) + 2\lambda_n J_1^2(\tilde{Q}^{\pi}) \\ &= 3\mathbb{P}_n\big[(1/T) \sum_{t=1}^{T} \hat{\mathcal{E}}_n^2(S_t, A_t; \eta^{\pi}, \tilde{Q}^{\pi})\big] + 2\lambda_n J_1^2(\tilde{Q}^{\pi}) \end{aligned}$$

$$= 3\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\hat{\mathcal{E}}_n(S_t,A_t;\eta^\pi,\tilde{Q}^\pi)-\mathcal{E}_{\mathcal{G}}(S_t,A_t;\eta^\pi,\tilde{Q}^\pi))^2\big]+2\lambda_n J_1^2(\tilde{Q}^\pi)$$
$$\le 3C(1+J_1^2(\tilde{Q}^\pi)+\log(1/\delta))\mu_n+2\lambda_n J_1^2(\tilde{Q}^\pi)\le c_0(1+\log(1/\delta))\lambda_n$$

For the second term,

$$I_{n,2}=(\mathbb{P}_n+P)f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))-3(\mathbb{P}_n f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))+(2/3)\lambda_n J_1^2(\hat{Q}_n))$$
$$=2(P-\mathbb{P}_n)f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))-Pf(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))-2\lambda_n J_1^2(\hat{Q}_n)$$

Recall from Lemma 3 that with probability $1-\delta/4$, for all (η,Q), $J_2(\hat{\mathcal{E}}_n(\eta,Q))\le c_0(J_1(Q)+J_2(\mathcal{E}_{\mathcal{G}}(\eta,Q))+\sqrt{\log(4/\delta)})$ for some constant c_0 when $\mu_n=O_P(n^{-1/(1+\alpha)})$. Combining with Assumption 5, for some constant c_1 (depending c_0 and C_1,C_2 in Assumption 5), we have $\Pr(E)>1-\delta/4$, where the event E is given by

$$E=\{J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))\le c_1(1+J_1(\hat{Q}_n)+\sqrt{\log(4/\delta)})\}$$

Now we have $\Pr(I_{n,2}>t)\le\Pr(I_{n,2}>t,E)+\delta/4$ and we bound the first term using peeling device on $2\lambda_n J_1^2(\hat{Q}_n)$ in $I_{n,2}$; similar as in the proof of Lemma 3. In particular,

$$\Pr(I_{n,2}>t,E)$$
$$=\sum_{l=0}^{\infty}\Pr\left(I_{n,2}>t,2\lambda_n J_1^2(\hat{Q}_n)\in[2^l t\mathbb{1}_{\{t\neq 0\}},2^{l+1}t),E\right)$$
$$=\sum_{l=0}^{\infty}\Pr\big(2(P-\mathbb{P}_n)f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))>Pf(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))+2\lambda_n J_1^2(\hat{Q}_n)+t,$$
$$2\lambda_n J_1^2(\hat{Q}_n)\in[2^l t\mathbb{1}_{\{t\neq 0\}},2^{l+1}t),J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))\le c_1(1+J_1(\hat{Q}_n)+\sqrt{\log(4/\delta)})\big)$$
$$\le\sum_{l=0}^{\infty}\Pr\big(2(P-\mathbb{P}_n)f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))>Pf(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))+2^l t\mathbb{1}_{\{t\neq 0\}}+t,$$
$$2\lambda_n J_1^2(\hat{Q}_n)\le 2^{l+1}t,J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))\le c_1(1+\sqrt{(2^l t)/\lambda_n}+\sqrt{\log(4/\delta)})\big)$$
$$\le\sum_{l=0}^{\infty}\Pr\big(2(P-\mathbb{P}_n)f(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))>Pf(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))+2^l t,$$
$$J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n))\le c_1(1+\sqrt{(2^l t)/\lambda_n}+\sqrt{\log(4/\delta)})\big)$$
$$\le\sum_{l=0}^{\infty}\Pr\left(\sup_{f\in\mathcal{F}_l}\frac{(P-\mathbb{P}_n)f(\mathcal{D})}{Pf(\mathcal{D})+2^l t}>\frac{1}{2}\right)$$

where $\mathcal{F}_l = \{f(g) : J_2(g) \le c_1(1 + \sqrt{(2^l t)/\lambda_n} + \sqrt{\log(4/\delta)}), g \in \mathcal{G}\}$. In what follows we verify the conditions in Theorem 19.3 in [74] with $\mathcal{F} = \mathcal{F}_l$, $\epsilon = 1/2$ and $\eta = 2^l t$ to get an exponential inequality for each term in the summation.

(A1) $|f(g)(\mathcal{D})| = |\frac{1}{T}\sum_{t=1}^{T} g(S_t, A_t)^2| \le G_{\max}^2$. We set $K_1 = G_{\max}^2$.

(A2) $Pf^2(g) \le G_{\max}^2 Pf(g)$. We set $K_2 = G_{\max}^2$

(A3) the condition $\sqrt{n}\epsilon\sqrt{1-\epsilon}\sqrt{\eta} \ge 288 \max\{2K_1, \sqrt{2K_2}\}$ becomes

$$\sqrt{n}(1/2)^{3/2} 2^{(l+1)/2}\sqrt{t} \ge 288 \max\{2G_{\max}^2, \sqrt{2}G_{\max}\}$$

So this holds for all $l \ge 0$ as long as $\sqrt{n}(1/2)^{3/2}\sqrt{2}\sqrt{t} \ge 288 \max\{2G_{\max}^2, \sqrt{2}G_{\max}\}$, i.e., $t \ge c_2/n$ for some constant c_2.

(A4) We first obtain an upper bound $\mathcal{N}_2(u, \mathcal{F}_l, \mathcal{D}_1, \ldots, \mathcal{D}_n)$ for all possible realization of trajectories. Firstly, for two g_1, g_2

$$\frac{1}{n}\sum_{i=1}^{n} [f(g_1)(\mathcal{D}_i) - f(g_2)(\mathcal{D}_i)]^2 \le 4G_{\max}^2 \|g_1 - g_2\|_{n,T}^2$$

where the norm $\|g\|_{n,T}^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} g^2(S_{i,t})$. Thus applying Assumption 6 implies that for some constant c_3, the metric entropy for each l is bounded by

$$\begin{aligned}
&\log \mathcal{N}_2(u, \mathcal{F}_l, \mathcal{D}_1, \ldots, \mathcal{D}_n) \\
&\le \log \mathcal{N}_2\left(\frac{u}{2G_{\max}}, \{g : J_2(g) \le c_1(1 + \sqrt{(2^l t)/\lambda_n} + \sqrt{\log(4/\delta)}), g \in \mathcal{G}\}, \{S_{i,t}, A_{i,t}\}_{i,t=1}^{N,T}\right) \\
&\le C\left(\frac{c_1(1 + \sqrt{(2^l t)/\lambda_n} + \sqrt{\log(4/\delta)})}{u/(2G_{\max})}\right)^{2\alpha} \le c_3\left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \log^{\alpha}(4/\delta)\right) u^{-2\alpha}
\end{aligned}$$

where C in the second last inequality is the constant specified in Assumption 6 Now we just need to ensure for all $x \ge \eta/8 = 2^l t/8$ and $l \ge 0$:

$$\frac{\sqrt{n}(1/2)^2 x}{96\sqrt{2} \cdot 2G_{\max}^2} \ge \int_0^{\sqrt{x}} \sqrt{c_3}\left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \log^{\alpha}(4/\delta)\right)^{1/2} u^{-\alpha} du$$

Note that $\int_0^{\sqrt{x}} u^{-\alpha} du = x^{\frac{1-\alpha}{2}}$. The above equality is equivalent with he following for some

constant c_4 (free of l, λ_n, n)

$$c_4\sqrt{n}x^{\frac{1+\alpha}{2}} \geq \left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \log^{\alpha}(4/\delta)\right)^{1/2}$$

Since $(a+b)^{1/2} \leq \sqrt{a} + \sqrt{b}$ and LHS is increasing function of x, it's enough to ensure

$$\begin{aligned} c_4\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq 1 \\ c_4\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq \left(\frac{2^l t}{\lambda_n}\right)^{\alpha/2} \\ c_4\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq (\log(4/\delta))^{\alpha/2} \end{aligned}$$

hold for all $l \geq 0$. The above is satisfied for all l by choosing large enough t. For example, the first one holds $t \geq c_5 n^{-1/(1+\alpha)}$ where $c_5 = 8^{(1+\alpha)}/c_4^2$. Similarly, we can verify that the second and third conditions holds if $t \geq \frac{c_5}{n\lambda_n^{\alpha}}$ and $t \geq c_5(\log(4/\delta))^{\alpha} n^{-1/(1+\alpha)}$. In summary, we select t s.t.

$$t \geq c_5(1 + (\log(4/\delta))^{\alpha})n^{-1/(1+\alpha)} + c_5(n\lambda_n^{\alpha})^{-1}$$

We can now apply Theorem 19.3 in [74] for each l-th term. For some constant c_6, c_7 depending on $G_{\max}$ we have

$$\Pr(I_{n,2} > t, E) \leq \sum_{l=0}^{\infty} \Pr\left(\sup_{f \in \mathcal{F}_l} \frac{(P - \mathbb{P}_n)f(\mathcal{D})}{Pf(\mathcal{D}) + 2^l t} > \frac{1}{2}\right) \leq c_6 \exp(-c_7 n t)$$

Solving $c_6 \exp(-c_7 n t) < \delta/4$, we get $t \geq \frac{\log(4c_6/\delta)}{nc_7}$. Combining this with the conditions on t in (A3) and (A4) implies that with probability $1 - \delta/2$,

$$I_{n,2} \leq \frac{c_2}{n} + \frac{c_5(1 + (\log(4/\delta))^{\alpha})}{n^{1/(1+\alpha)}} + \frac{c_5}{n\lambda_n^{\alpha}} + \frac{\log(4c_6/\delta)}{nc_7}$$

With the choice of tuning parameters (λ_n, μ_n) specified in the condition , we have $(n\lambda_n^{\alpha})^{-1}$ and $(1/n)^{1/(1+\alpha)} = O_P(\lambda_n)$ and thus $I_{n,2} \leq c(\delta)\lambda_n$ for some constant $c(\delta)$. Combining with the bound on $I_{n,1}$, we obtain the desired result. □

Lemma 5. *Suppose Assumption 1 holds. Then, for all* $(\eta, Q) \in \mathbb{R} \times \mathcal{Q}$*, we have* $|\eta - \eta^{\pi}| \lesssim \|\mathcal{E}(\eta, Q)\|_{\bar{\nu}_T}$ *and* $\|Q - \tilde{Q}^{\pi}\| \lesssim \|\mathcal{E}(\eta, Q)\|_{\bar{\nu}_T}$

Proof of Lemma 5. Note that

$$\begin{aligned}
&\mathcal{E}(s,a;\eta,Q) = \mathbb{E}[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1},a') - \eta - Q(s,a) \mid S_t = s, A_t = a] \\
&= (\eta^\pi - \eta) + (\tilde{Q}^\pi - Q)(s,a) - \mathcal{P}^\pi(\tilde{Q}^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + (\eta^\pi - \eta)u^\pi(s,a) + (\tilde{Q}^\pi - Q)(s,a) - \mathcal{P}^\pi(\tilde{Q}^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + (\eta^\pi - \eta)(q^\pi(s,a) - \mathcal{P}^\pi q^\pi(s,a)) + (\tilde{Q}^\pi - Q)(s,a) - \mathcal{P}^\pi(\tilde{Q}^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + h(s,a) - \mathcal{P}^\pi h(s,a)
\end{aligned}$$

where $h = \tilde{Q}^\pi - Q + (\eta^\pi - \eta)q^\pi$. Using (5.10), we have

$$\|\mathcal{E}(\eta,Q)\|_{\bar{\nu}_T}^2 = (\eta - \eta^\pi)^2\|e^\pi\|_{\bar{\nu}_T}^2 + \|h\|_{\bar{\nu}_T}^2$$

We have $|\eta - \eta^\pi| \le \|e^\pi\|_{\bar{\nu}_T}^{-1}\|\mathcal{E}(\eta,Q)\|$. On the other hand, we have

$$\begin{aligned}
&\|(\mathcal{I} - \mathcal{P}^\pi)(Q - Q^\pi)\|_{\bar{\nu}_T} \\
&= \|Q - \mathcal{P}^\pi Q + \eta\mathbf{1} - r + (r + \mathcal{P}^\pi Q^\pi - Q^\pi - \mathbf{1}\eta^\pi) - \eta\mathbf{1} + \eta^\pi\mathbf{1}\|_{\bar{\nu}_T} \\
&= \| - \mathcal{E}(\eta,Q) - \eta\mathbf{1} + \eta^\pi\mathbf{1}\|_{\bar{\nu}_T} \\
&\le \|\mathcal{E}(\eta,Q)\|_{\bar{\nu}_T} + |\eta^\pi - \eta| \le (1 + \|e^\pi\|_{\bar{\nu}_T}^{-1})\|\mathcal{E}(\eta,Q)\|_{\bar{\nu}_T}
\end{aligned}$$

□

Proof of Theorem 2. Let the objective function in (5.7) be $L_n(\eta, Q)$:

$$L_n(\eta,Q) = \mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{\mathcal{E}}_n^2(S_t,A_t;\eta,Q)\big] + \lambda_n J_1^2(Q)$$

Recall the definition of $\tilde{q}^\pi$ in Assumption 7. Note that we assume $\tilde{q}^\pi \in \mathcal{Q}$. Consider

$$\begin{aligned}
&\frac{d}{dt}L_n(\eta - t, Q + t\tilde{q}^\pi)|_{t=0} \\
&= \frac{d}{dt}\mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{\mathcal{E}}_n^2(S_t,A_t;\eta - t, Q + t\tilde{q}^\pi)\big] + \lambda_n J_1^2(Q + t\tilde{q}^\pi)|_{t=0} \\
&= 2\mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{\mathcal{E}}_n(S_t,A_t;\eta,Q) \times \frac{d}{dt}\hat{\mathcal{E}}_n(S_t,A_t;\eta - t, Q + t\tilde{q}^\pi)|_{t=0}\big] + 2\lambda_n J_1(Q,\tilde{q}^\pi)
\end{aligned}$$

Using the optimizing property (5.8), we can show that $\frac{d}{dt}\hat{\mathcal{E}}_n(s,a;\eta-t,Q+t\tilde{q}^\pi)|_{t=0} = \hat{e}_n(s,a)$ where $\hat{e}_n$ solves

$$\hat{e}_n = \operatorname*{argmin}_{g\in\mathcal{G}} \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}(1-\tilde{q}^\pi(S,A)+\sum_{a'}\pi(a'|S')\tilde{q}^\pi(S',a')-g(S_t,A_t))^2\big] + \mu_n J_2^2(g) \quad (5.16)$$

See the derivation below. From (5.11), $\hat{e}_n$ can be viewed as an estimator of $e^\pi = 1-u^\pi$. Since $(\hat{\eta}_n,\hat{Q}_n) = \operatorname{argmin}_{\eta,Q} L_n(\eta,Q)$, we have $\frac{d}{dt}L_n(\hat{\eta}_n - t, \hat{Q}_n + t\tilde{q}^\pi)\big|_{t=0} = 0$, that is

$$0 = \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)\hat{e}_n(S_t,A_t)\big] + \lambda_n J_1(\hat{Q}_n,\tilde{q}^\pi) \quad (5.17)$$

From Theorem 1, we have $J_1(\hat{Q}_n) = O_P(1)$ and thus the second term $\lambda_n J_1(\hat{Q}_n,\tilde{q}^\pi) = O_P(n^{-1/2})$ since $\lambda_n = o_P(n^{-1/2})$. For the first term,

$$\begin{aligned}
&\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)\hat{e}_n(S_t,A_t)\big] \\
&= \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)e^\pi(S_t,A_t)\big] + \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)(\hat{e}_n-e^\pi)(S_t,A_t)\big]
\end{aligned}$$

Since $e^\pi \in \mathcal{G}$ and the tuning parameter satisfies $\mu_n^{-1} = O_P(n^{-1/(1+\alpha)})$, applying the same argument as in Lemma 3 implies that $\|\hat{e}_n - e^\pi\|_n^2 = O_P(\mu_n)$. In Lemma 4, we have shown $\|\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n)\|_n^2 = O_P(\lambda_n)$. Note $\mu_n = O_P(\lambda_n)$ and $\lambda_n = o_P(n^{-1/2})$, so that $\mu_n = o_P(n^{-1/2})$. By Cauchy inequality, $\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)(\hat{e}_n-e^\pi)(S_t,A_t)\big] = o_P(n^{-1/2})$.

On the other hand, since $e^\pi \in \mathcal{G}$, the optimizing property in (5.8) gives

$$\begin{aligned}
&\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\hat{\mathcal{E}}_n(S_t,A_t;\hat{\eta}_n,\hat{Q}_n)e^\pi(S_t,A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\delta_t(\hat{\eta}_n,\hat{Q}_n)e^\pi(S_t,A_t)\big] - 2\mu_n J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n),e^\pi)
\end{aligned}$$

where we use the notation for the temporal difference error at time t, i.e.,

$$\delta_t(\eta,Q) = \delta(S_t,A_t,S_{t+1},R_{t+1};\eta,Q)$$

Note that Lemma 3 implies that $J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n,\hat{Q}_n)) = O_P(1)$ since $\mu_n^{-1} = O_P(n^{1/(1+\alpha)})$. Thus the

second term of above $\mu_n J_2(\hat{\mathcal{E}}_n(\hat{\eta}_n, \hat{Q}_n), e^\pi) = o_P(n^{-1/2})$. Finally we consider the first term. Plugging-in the the true tempera difference error, we have

$$\begin{aligned}
&\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\delta_t(\hat{\eta}_n, \hat{Q}_n)e^\pi(S_t, A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\delta_t(\eta^\pi, \tilde{Q}^\pi) + (\delta_t(\hat{\eta}_n, \hat{Q}_n) - \delta_t(\eta^\pi, \tilde{Q}^\pi))e^\pi(S_t, A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\delta_t(\eta^\pi, Q^\pi)e^\pi(S_t, A_t)\big] - (\hat{\eta}_n - \eta^\pi)\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}e^\pi(S_t, A_t)\big] + \text{Rem}
\end{aligned}$$

where the last term is given by

$$\text{Rem} = \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}[(\tilde{Q}^\pi - \hat{Q}_n)(S_t, A_t) - \sum_{a'}\pi(a'|S_{t+1})(\tilde{Q}^\pi - \hat{Q}_n)(S_{t+1}, a')]e^\pi(S_t, A_t)\big]$$

For $Q \in \mathcal{Q}$, define

$$f(Q) : \mathcal{D} \to \frac{1}{T}\sum_{t=1}^{T}[Q(S_t, A_t) - \sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1}, a')]e^\pi(S_t, A_t)$$

Note that we can write $Pf(Q) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[h(S_t, A_t)e^\pi(S_t, A_t)]$ where $h(s, a) = Q(s, a) - \sum_{a'}\mathbb{E}[\pi(a'|S_{t+1})Q(S_{t+1}, a')|S_t = a, A_t = a]$ is mean 0 under stationary distribution. Using the orthogonality (5.10), we have $Pf(Q) = 0$ and thus Rem $= -\sqrt{n}(\mathbb{G}_n f(\hat{Q}_n) - \mathbb{G}_n f(\tilde{Q}^\pi))$. Note that $J_1(\hat{Q}_n) = O_P(1)$ and the sup-norm metric condition (6) implies that the bracket entropy integral of the function class $\mathcal{F} = \{f(Q) : Q \in \mathcal{Q}, J(Q) \leq M\}$ is finite for all M. Using Assumptions 3 and 7 and Lemma 5, we have

$$\begin{aligned}
P(f(\hat{Q}_n) - f(\tilde{Q}^\pi))^2 &\leq G_{\max}^2(1 + 4Q_{\max}^2)\|(\mathcal{I} - \mathcal{P}^\pi)(\hat{Q}_n - \tilde{Q}^\pi)\|_{\bar{\nu}_T}^2 \\
&\leq G_{\max}^2(1 + 4Q_{\max}^2)(1 + \|e^\pi\|_{\bar{\nu}_T}^{-1})\|\mathcal{E}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T}
\end{aligned}$$

where $\mathcal{I}$ is the identity operator and recall that $\mathcal{P}^\pi Q(s, a) = \mathbb{E}[\sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1}, a')|S_t = a, A_t = a]$. Since $\|\mathcal{E}(\hat{\eta}_n, \hat{Q}_n)\|_{\bar{\nu}_T} = O_P(\lambda_n) = o_P(1)$, we conclude that $P(f(\hat{Q}_n) - f(\tilde{Q}^\pi))^2 = o_P(1)$. The asymptotic equicontinuity (Theorem 19.5 in [71]) implies that $|\mathbb{G}_n f(\hat{Q}_n) - \mathbb{G}_n f(\tilde{Q}^\pi)| = o_P(1)$ and thus Rem $= o_P(n^{-1/2})$.

Summarizing above and plugging into (5.17), we have shown that

$$(\hat{\eta}_n - \eta^\pi)\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} e^\pi(S_t, A_t)\big] = \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \delta_t(\eta^\pi, Q^\pi)e^\pi(S_t, A_t)\big] + o_P(n^{-1/2})$$

Note that $\mathbb{P}_n(1/T)\sum_{t=1}^{T} e^\pi(S_t, A_t) = 1/\tau^\pi + o_P(1)$ and $w^\pi = \tau^\pi e^\pi$. We then have

$$\sqrt{n}(\hat{\eta}_n - \eta^\pi) = \mathbb{G}_n\big[(1/T)\sum_{t=1}^{T} \delta_t(S_t, A_t; \eta^\pi, Q^\pi)w^\pi(S_t, A_t)\big] + o_P(1)$$

□

Derivation of Equation (5.16) . Note that for all $g \in \mathcal{G}$, $\hat{\mathcal{E}}_n(\eta, Q)$ satisfies

$$\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\Big(\delta(S_t, A_t, S_{t+1}, R_{t+1}; \eta, Q) - \hat{\mathcal{E}}_n(S_t, A_t; \eta, Q)\Big) g(S_t, A_t)\big] = \mu_n J_2(g, \hat{\mathcal{E}}_n(\eta, Q))$$

Now for all t and $g \in \mathcal{G}$, we have

$$\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}[\delta(S_t, A_t, S_{t+1}, R_{t+1}; \hat{\eta}_n - t, \hat{Q}_n + t\tilde{q}^\pi) - \hat{\mathcal{E}}_n(S_t, A_t; \hat{\eta}_n - t, \hat{Q}_n + t\tilde{q}^\pi)]g(S_t, A_t)\big]$$
$$= \mu_n J_2(g, \hat{\mathcal{E}}_n(\hat{\eta}_n - t, \hat{Q}_n + t\tilde{q}^\pi))$$

Taking derivative w.r.t. t at $t = 0$ of above gives

$$\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}[1 - \tilde{q}^\pi(S_t, A_t) + \sum_{a'}\pi(a'|S_{t+1})\tilde{q}^\pi(S_{t+1}, a') - \hat{e}_n(S_t, A_t)]g(S_t, A_t)\big] = \mu_n J_2(g, \hat{e}_n)$$

where we use the fact that

$$\frac{d}{dt}\delta(S, A, S', R; \hat{\eta}_n - t, \hat{Q}_n + t\tilde{q}^\pi) = 1 - \tilde{q}^\pi(S, A) + \sum_{a'}\pi(a'|S')\tilde{q}^\pi(S', a')$$

And thus we can see that $\hat{e}_n$ solves

$$\hat{e}_n = \underset{g\in\mathcal{G}}{\operatorname{argmin}}\,\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}[1 - \tilde{q}^\pi(S_t, A_t) + \sum_{a'}\pi(a'|S_{+1})\tilde{q}^\pi(S_{t+1}, a') - g(S_t, A_t)]^2\big] + \mu_n J_2^2(g)$$

□

CHAPTER 6

Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity

6.1 Introduction

With the recent evolution of mobile health technologies, health scientists are increasingly interested in delivering interventions via notifications on the mobile device at the moments when they can most readily help the user prevent negative health outcomes and promote the adoption and maintenance of healthy behaviors. The type and timing of the mobile health interventions should ideally adapt to the real-time collected user's context, e.g., the time of the day, the location, current activity, and stress level. This gives rise to the concept of a just-in-time adaptive intervention (JITAI) [36]. Operationally, JITAI includes a sequence of decision rules (e.g., treatment policies) that takes the user's current context as input and specifies whether and what type of an intervention should be provided at the moment. In practice, the interventions included in a JITAI are often based on behavioral theories. However, these theories are currently not mature enough to precisely specify which particular intervention and when it should be delivered in order to ensure the interventions have the intended effects and optimize the long-term efficacy of the interventions. As a result, there is much interest in how best to use data to design JITAIs.

In this chapter, we develop a Reinforcement Learning (RL) algorithm to continuously learn and improve the treatment policy in the JITAI as the data is collected from the user. This work is motivated by our collaboration on a physical activity mobile health study, called HeartSteps V2. The HeartSteps V2 RL algorithm will be used to decide, five times per day, whether to deliver a context-tailored activity suggestion.

The remainder of the chapter is organized as follows. We first describe HeartSteps, including HeartSteps V1, and the future planned study, HeartSteps V2. We then briefly discuss some RL background and identify key challenges in applying RL to optimize JITAI treatment policies in

mobile health. Existing mobile health studies that utilized RL are reviewed, as well as related RL algorithms. We then describe the proposed HeartSteps V2 RL algorithm, the implementation, and evaluation of this algorithm using a generative model built on HeartSteps V1 data. We close with a discussion of future work.

6.2 HeartSteps V1 and V2: Physical Activity Mobile Health Study

In the upcoming HeartSteps V2, a 90-day physical activity study, patients with blood pressure in the stage 1 hypertension range (120-130 systolic) will be provided a Fitbit tracker and a mobile phone application on the phone designed to help them improve their physical activity. One of the interventions in HeartSteps V2 is the contextually tailored activity suggestion that may be delivered at any of the five user-specified times during each day. These five times are roughly separated by 2.5 hours, corresponding to the user's morning commute, mid-day, mid-afternoon, evening commute, and post-dinner times. The content of the suggestion is designed to encourage activity in the current context and thus the suggestions are intended to impact near time physical activity. The RL algorithm developed in this chapter will be used to determine at each time whether to send the activity suggestion.

In order to design HeartSteps V2, our team conducted HeartSteps V1, which is a 42-day physical activity study involving 37 healthy sedentary adults [1, 47, 48, 49]. In HeartSteps V1 whether to provide a tailored activity suggestion were randomized at each of the 5 times per day with a constant probability of 0.30. The data collected from HeartSteps V1 is used in this chapter to (1) inform the design of RL algorithm for HeartSteps V2 (e.g., selecting the variables that are predictive of future step counts as well as the efficacy of the activity suggestion and form a prior distribution) and (2) to create a simulation environment (e.g., the generative model) in order to design/evaluate the RL algorithm. See sections 6.5 and 6.6.

6.3 Challenges to Applying Reinforcement Learning in mHealth

Reinforcement Learning (RL) is an area of Machine Learning in which an algorithm learns how to act optimally by continuously interacting with the unknown environment: observe the current state, perform the action and receive the reward, with the goal of learning the best sequence of actions (i.e. the policy) to maximize the total rewards. For example, in the case of HeartSteps, the

state is a set of features of the user's current and past context, the actions are whether to deliver an activity suggestion and the reward is a function of near time physical activity. The fundamental challenge in RL is the trade-off between exploitation (e.g., taking the action that appears the best given data observed so far) and exploration (e.g., gathering information to infer the best action). RL has seen rapid development in recent years and shown remarkable success across many fields, e.g., video games, chess-playing, and robotic control. However, many challenges remain that need to be carefully addressed before RL can be usefully deployed to adapt and optimize mobile health interventions. Below we discuss some of these challenges.

First, the RL algorithm should learn quickly. Most online RL algorithms require the agent to interact many times with the environment prior to performing well. This is impractical in mobile health applications as users can lose interest and disengage quickly. We want to ensure that the RL algorithm performs well sufficiently fast.

Second, the RL algorithm must accommodate noisy data. Because mobile health interventions are provided in uncontrolled, in situ complex environments both context information, as well as rewards, can be very noisy. For example, step count data collected from the wrist band is noisy due to a variety of confounds including incidental hand movements. Additionally, the sensors do not detect the entire context of the user; non-sensed aspects of the current context act as sources of variance. Such high noise settings typically require even more interactions with the environment to learn the optimal policy. Both the first and second challenges result in the need to trade off between bias and variance when designing the RL algorithm.

Third, the RL algorithm should accommodate some non-stationarity. Due to unobserved aspects of the current context (e.g., engagement or burden), observed human behavior often exhibits non-stationarity over longer periods of time, e.g., the mapping from context to reward will likely change slowly. For example, in HeartSteps V1 treatment effects decrease with the time the user is in the study[47].

Fourth, the RL algorithm must adjust for longer-term effects of current actions. In mobile health interventions often tend to have a positive effect on the immediate reward, but likely produce a negative impact on future rewards due to user habituation and/or burden [45, 46]. As such, the optimal treatment policy can only be identified by looking far into the future, e.g., using a long planning horizon in RL. It's been shown that, in both practice and theory, larger discount rates lead to slower learning rates.

Lastly, the RL algorithm should select actions so that after the study is over, secondary data analyses are feasible. This is particularly the case for experimental trials involving clinical populations. In these settings, an interdisciplinary team is required to design the intervention and to

conduct the experimental trial. As a result, multiple stakeholders will want to analyze the resulting data in a large variety of ways. Thus, for example, off-policy learning and causal inference, as well as other more standard statistical analyses, must be feasible after study end.

6.4 Related Work

Existing RL-based Mobile Health Studies There are a few existing mobile health studies in which RL methods are applied to adapt the individual's intervention in real time. Here we only focus on the setting where the treatment policy is not pre-specified, but instead continuously learned and improved as more data is collected.

In [76], an RL system was deployed to choose the different types of daily suggestions to encourage physical activity in patients with diabetes in a 26-week study. They used the Softmax strategy to optimize the intervention in the so-called Contextual Bandit setting, e.g., the action (daily suggestion) is chosen with the goal to maximize the immediate reward (increased step count). Paredes et al. [77] developed PopTherapy, an RL system to choose among 10 types of stress management strategies when the participant requests an intervention in the mobile app. A Contextual Bandit algorithm, called Upper Confidence Bound (UCB), was applied in their RL system. A recent weight loss study is reported in [78], in which one of three types of interventions is chosen twice a week over a 12-week period. Their RL system features an explicit separation between exploration and exploitation, e.g., 10 decision times are predetermined for exploration (e.g., randomly selecting the interventions at each decision time) and the rest of 14 decision times for exploitation (e.g., choosing the best intervention that maximizes the designed reward based on the history). In MyBehavior [79], a smartphone app that delivered personalized interventions for promoting physical activity and dietary health, used EXP3, a multi-arm bandit algorithm (e.g., context-free) to select the interventions. While the RL methods in the aforementioned studies aim to select actions so as to optimize the immediate reward, in a recent physical study reported in [80], the RL system at the end of every week uses the participant's historical daily step count data to estimate the dynamic system for the daily step count and use it to infer the optimal daily step goals for the next 7 days, with the goal to maximize the minimal number of step counts taken in the next week.

The challenges listed in section 6.3 motivate us to generalize the above RL algorithms in particular directions. First, the use of a Bayesian RL algorithm will allow us to include prior evidence to accelerate learning on each user. Second, we will trade off bias with variance by using a low dimensional linear model for the part of the reward function corresponding to the treatment effect. Third, we will use a modeling approach that increases the robustness of the RL algorithm

to reward function misspecification as it is unlikely that the model used to approximate the reward function will be correct (due to the dimension and complexity of the context information and potential non-stationarity). It is been empirically shown in RL literature that the performance of standard RL algorithms are quite sensitive to the model for the reward function [81, 82, 83]. Third, the RL algorithm should accommodate delayed habituation and burden effects. Among the above algorithms, only the algorithm used in [80] attempts to optimize rewards over a time period longer than the immediate time step. It turns out that there is a bias-variance trade-off when designing how long into the future the RL should attempt to optimize rewards. That is, only focusing on maximizing the immediate rewards speeds the learning rate (e.g., due to lower estimation variance) compared with a full RL algorithm that attempts to maximize over a longer time horizon. However, an RL algorithm focused on optimizing the immediate reward might end up sending too many treatments due to the fourth mentioned challenge (i.e. the treatment tends to have a positive effect on immediate reward and negative effects on future rewards) and lead to poorer overall performance (akin to bias) than the algorithm that attempts to optimize over a longer time horizon to account for treatment burden and disengagement. Lastly, it is critical that the RL algorithm select actions with probabilities bounded away from 0 and 1 to ensure off-policy analyses and minimize assumptions needed to conduct causal inference. Both [77] and [80] use algorithms that select the action deterministically based on the history, and [78] incorporate a pure exploitation phase. It's known that action selection probabilities close to 0 or 1 cause the instability (i.e., high variance) in batch data analysis that uses importance weights, e.g., in the off-policy evaluation [84, 85]

Related RL Algorithms At a high level, our proposed algorithm includes three main components: (1) actions are selected probabilistically based on Thompson Sampling (TS), (2) action-centering is used in the model for the reward function and (3) a proxy for future rewards is used to capture the potential negative future impact of sending treatments. Below we briefly discuss how each of these components is related to existing RL algorithms.

Thompson Sampling (TS) is a general algorithmic idea that uses a Bayesian paradigm to trade off between exploration and exploitation. In the standard TS algorithm, the action at each decision time is selected according to the posterior probability of this action being optimal given the current history. TS-based algorithms have been shown to enjoy not only strong theoretical performance guarantees but strong empirical performance in many problems when compared to other state-of-the-art methods, such as UCB [86, 87, 88]. In addition, prior knowledge can be easily incorporated in the TS based algorithm, through the use of a prior distribution on the parameters.

We use the idea of action-centering in modeling the reward. The motivation is to protect the

algorithm from a misspecified model for the "baseline" reward function (e.g., in HeartSteps example with binary actions, the baseline reward function is the expected number of future 30-min step count given the current state and no activity suggestion). The idea of action-centering in RL was first explored in [89] and recently improved in [90]. In both works, the RL algorithm is theoretically guaranteed to learn the optimal action under no assumption about the baseline reward generating process (e.g., the baseline reward function can be non-stationary). However, neither of these methods attempts to reduce the noise in the reward. We generalize the action centering for use in higher variance, non-stationary reward settings.

To capture the potential negative impact of treatment on future rewards, we introduce a "dosage" variable based on the history of past treatments. A similar formulation of "dosage" was explored in a recent unpublished manuscript [83] where they developed a bandit algorithm, called ROGUE (Reducing or Gaining Unknown Efficacy) Bandits. They use the "dosage" idea to accommodate settings in which an (unknown) dosage variable causes non-stationarity in the reward function. Our use of dosage, on the other hand, is to form a proxy of the future rewards, in order to mimic a full RL setting (as opposed to the bandit setting) but managing variance. We construct a proxy of the future rewards (proxy value) under a low dimensional proxy MDP model. The model-based RL approach is well studied in the RL literature, for example in TS-based algorithms in [91, 92, 93]. In these work, the algorithm uses a model for the transition function from current state and action to the next state. Instead, the proposed algorithm in this chapter only uses the proxy MDP model to provide a low variance proxy for the longer term impact of actions on future rewards.

6.5 Reinforcement Learning Algorithm in HeartSteps V2

In this section, we discuss the design of the RL algorithm in HeartStep V2; this algorithm determines whether to send the activity suggestion at each decision time. We first detail the underlying RL framework by specifying each component in our setting, i.e. the decision time, action, states, and reward, and formally introduce our proposed RL algorithm.

Reinforcement Learning Framework

Let the participant's longitudinal data recorded via mobile device be the sequence

$$\{S_1, A_1, R_1, S_2, A_2, R_2, \ldots, S_t, A_t, R_t, \cdots\}$$

Here t indexes decision time. In HeartSteps V1, as in the planned HeartSteps V2, there are five decision times each day. We also use (l, d) to refer the l-th time decision time on study day d. For example, $(l, d) = (5, 3)$ refers to the 5-th time in day 3, which corresponds to time $t = 5(d-1) + l = 15$. A_t is the action or treatment at time t. The treatment is binary, i.e. $A_t = 1$ if an activity suggestion is delivered and $A_t = 0$ otherwise. R_t is the immediate reward collected after action A_t. In HeartSteps, the reward is the log transformation of the step count collected 30 minutes after the decision time. S_t is the state vector at decision time t. We decompose the state vector as $S_t = \{I_t, Z_t, X_t\}$. I_t is used to indicate times at which only $A_t = 0$ is feasible and/or ethical. For example, if sensors indicate that the participant might be driving a car, then the suggestion should not be sent; that is, the participant is unavailable for treatment ($I_t = 0$). Z_t denotes features used to represent the current context at time t. In HeartSteps, these features include current location, the prior 30-minute step count, yesterday's daily step count, the current temperature, as well as the measures of how active the participant has been around the current decision time over the last week. Lastly, X_t is the "dosage" variable that captures our proxy for the treatment burden, defined based on the participant's treatment history. In contrast to HeartSteps V1, in HeartSteps V2, an additional intervention component, i.e., an anti-sedentary suggestion will sometimes be delivered when the participant is sedentary. As the anti-sedentary suggestion, in addition to the activity suggestions, can cause the burden, it is included in defining the dosage variable. Specifically, denote by E_t the event that an activity suggestion is sent at decision time $t-1$ (e.g., $A_{t-1} = 0$) and any anti-sedentary suggestion is sent between time $t-1$ and t. The dosage at the moment is constructed by first multiplying the previous dosage variable by $\lambda \in (0, 1)$ and incrementing it by 1 if any messages were sent to the user since last decision time. Specifically, starting with the initial value $X_1 = 0$, the dosage at time $t+1$ is defined as $X_{t+1} = \lambda X_t + \mathbb{1}_{E_{t+1}}$. Based on the data analysis result from HeartSteps V1, we choose $\lambda = 0.9$; see section 6.5 for how this value is selected.

At each decision time, the RL algorithm selects the action based on each participant's current history (e.g., the past states, actions, and rewards), with the goal to optimize the total rewards during the process. The proposed algorithm is stochastic, that is, the algorithm will output a probability to select an action. Denote the history up to the end of day d by $H_d = \{S_{t,k}, A_{t,k}, R_{t,k}\}_{1 \le t \le 5, 1 \le k \le d}$. The RL algorithm consists two components: (1) the nightly update, e.g., $H_{d-1} \mapsto \{(\mu_d, \Sigma_d), \eta_d\}$ where (μ_d, Σ_d) denote parameters in the posterior distribution for the reward and η_d proxies the delayed effect on future rewards, both calculated at the end of the previous day (see below for more details), and (2) the probability $\pi_{l,d}$, to select the action (e.g., $A_{l,d}$ is sampled from a Bernoulli distribution with probability $\pi_{l,d}$). See Figure 3 for the pseudo code of the proposed HeartSteps V2

RL algorithm.

ALGORITHM 3: HeartSteps V2 RL Algorithm

Input: feature vectors $f(s)$ and $g(s)$, prior distributions $\boldsymbol{N}(\mu_{\alpha_0}, \Sigma_{\alpha_0})$ and $\boldsymbol{N}(\mu_\beta, \Sigma_\beta)$, variance of noise σ^2. discount rate λ in dosage, discount rate in proxy value γ, updating weight in proxy value w, weight of delayed effect in action selection ξ, clipped probability ϵ_0 and ϵ_1.

Initialize $X_{1,1} \leftarrow 1$, $\mu_1 \leftarrow \mu_\beta$, $\Sigma_1 \leftarrow \Sigma_\beta$

for *day* $d = 1, 2, \ldots, 90$ **do**

  for *time slot* $l = 1, 2, \ldots, 5$ **do**

    Check the participant's availability $I_{l,d}$

    Check event $E_{l,d}$ and calculate $X_{l,d}$ based on the previous dosage and event $E_{l,d}$

    Observe the context variable $Z_{l,d}$

    Form the state, $S_{l,d} = \{I_{l,d}, Z_{l,d}, X_{l,d}\}$

    if *available (*$I_{l,d} = 1$*)* **then**

      Calculate $\pi_{l,d}$ (6.2), based on $\{(\mu_d, \Sigma_d), \eta_d\}$

      Sample $A_{l,d}$ from a Bernoulli distribution with probability $\pi_{l,d}$

      Send the activity message if $A_{l,d} = 1$. Otherwise, do nothing

    end

    else

      Do nothing

    end

  end

  Calculate the posterior distribution μ_{d+1}, Σ_{d+1} based on the model (6.3)

  Calculate the proxy delayed effect η_{d+1} in 6.5

end

Action Selection

The reward function is given by $r_t(s, a) = \mathbb{E}[R_t \,|\, S_t = s, A_t = a, I_t = 1]$. The action selection developed here is based on a linear working model for the treatment effect:

$$r_t(s, 1) - r_t(s, 0) = f(s)^\top \beta \tag{6.1}$$

where the feature vector, $f(s)$, is selected based on the domain science as well as on data analyses using HeartSteps V1; see section 6.5 for the discussion of how the features are selected. At the l-th decison time on day d, availability is ascertained (i.e. $I_{l,d} = 1$). Then for $S_{l,d} = s$ and $X_{l,d} = x$, the action, $a = 1$ is selected with probability

$$\Pr\left\{f(s)^\top \beta > \xi \cdot \eta_d(x);\ \beta \sim \boldsymbol{N}(\mu_d, \Sigma_d)\right\}$$

where the random variable β, follows a Normal distribution $\boldsymbol{N}(\mu_d, \Sigma_d)$, e.g., the posterior distribution of the parameters, obtained at the end of previous day. The term $\eta_d(X_{l,d})$ proxies the long-term, negative effect of delivering the activity suggestion at the moment given the current dosage level $X_{l,d}$ (see the detailed formulation of η_d in section 6.5) and $\xi > 0$ controls the relative importance of maximizing the future rewards in the current action selection, compared to the immediate rewards. Note that when $\xi = 0$, we recover the bandit formulation, e.g., the action is selected to maximize the immediate rewards, ignoring any impact in the future. When $\xi > 0$, the action selection aims to balance between the immediate effect and the delayed effect. The selection of ξ is discussed in section 6.5. The probability of sending an activity suggestion, $\pi_{l,d}$ (for $I_t = 1$, $S_{l,d} = s$, $X_{l,d} = x$) is a clipped version of the above:

$$\pi_{l,d} = \phi\left(\Pr\left\{f(s)^\top \beta > \xi \cdot \eta_d(x);\ \beta \sim \boldsymbol{N}(\mu_d, \Sigma_d)\right\}\right). \tag{6.2}$$

The clipping function is $\phi(\pi) = \min(1 - \epsilon_0, \max(\pi, \epsilon_1)) \in [\epsilon_1, 1 - \epsilon_0]$ This restricts the randomization probability of sending nothing and of sending an activity suggestion to be at least ϵ_0 and ϵ_1, respectively. The probability clipping enables off-policy data analyses after the study is over and, furthermore, ensures that the RL algorithm will continue to explore and learn, instead of locking itself into a particular policy. In HeartSteps V2, $\epsilon_0 = 0.2$ and $\epsilon_1 = 0.1$.

Nightly Updates

The posterior distribution of β for the immediate treatment effect and the proxy for the delayed effect are updated at the end of each day. Operationally, the nightly update is a mapping: $H_d \mapsto \{(\mu_{d+1}, \Sigma_{d+1}), \eta_{d+1}\}$, that takes the current history up to day d as the input and outputs the posterior distribution and proxy of delayed effect, which are used in the action selection in the following day (i.e. during day $d + 1$). We discuss each of them in turn.

Posterior Update of Immediate Treatment Effect We use the following linear Bayesian regression "working model" for the reward to derive the posterior distribution for the treatment effect:

$$R_t = g(S_t)^\top \alpha_0 + \pi_t f(S_t)^\top \alpha_1 + (A_t - \pi_t) f(S_t)^\top \beta + \boldsymbol{N}(0, \sigma^2), \text{ if } I_t = 1 \tag{6.3}$$

so that the working model for the baseline reward (i.e., $a = 0$)

$$r_t(s, 0) = g(s)^\top \alpha. \tag{6.4}$$

The baseline feature vector $g(s)$ is selected based on the domain science and data analyses using HeartSteps V1; see section 6.5 for a discussion. The use of π_t in (6.3) is unusual but provides a number of advantages as follows. Consider the action-centered term, $(A_t - \pi_t)$, in the working model (6.3). As long as the treatment effect model (6.1) is correctly specified, the estimator of β based on the model (6.3) is guaranteed to be unbiased even when the baseline reward model (6.4) is incorrect [40], for example, due to the non-linearity in $g(s)$ or non-stationarity (α changes over time). That is, through the use of action centering, we achieve the robustness against mis-specification of the approximate baseline model, (6.4). The rationale of including the term $\pi_t f(S_t)$ in the Bayesian regression working model (6.3) is to capture the time-varying aspect of the main effect due to the action-centered term (e.g., π_t is continuously updated during the study). Omitting this term would reduce the number of parameters in the model but we have found that in experiments the inclusion of $\pi_t f(S_t)$ reduces the variance of the treatment effect estimates and thus speed the learning. Second, in the case where the treatment effect model (6.1) is incorrect, for example, the treatment effect is non-linear in $f(S_t)$ or is time non-stationary (with time-varying β), it can be shown [40] that the Bayesian regression provides a linear approximation to the treatment effect. When the action is not centered, the treatment effect estimates may not converge to any useful approximation at all, which could lead to bad performance in selecting the action.

The Bayesian model requires prior distributions on α_0, α_1 and β. Here the priors are independent and given by $\alpha_0 \sim \boldsymbol{N}(\mu_{\alpha_0}, \Sigma_{\alpha_0})$, $\alpha_1 \sim \boldsymbol{N}(\mu_\beta, \Sigma_\beta)$, $\beta \sim \boldsymbol{N}(\mu_\beta, \Sigma_\beta)$. Because the priors are Gaussian and the error in (6.3) is Gaussian, the posterior distribution of β given the current history H_d is also a Gaussian, denoted by $\boldsymbol{N}(\mu_{d+1}, \Sigma_{d+1})$. See in section 6.5 for a discussion of how the prior is constructed using HeartStep V1 data.

Proxy Delayed Effect on Future Rewards The proxy is formed based on a simple Markov Decision Process (MDP) for the states $S_t = (Z_t, I_t, X_t)$, in which we make the following working assumptions:

(S1) the context $\{Z_t\}$ is i.i.d. with distribution F,

(S2) the availability $\{I_t\}$ is i.i.d. with probability p_{avail}

(S3) the dosage variable $\{X_t\}$ makes transitions according to $\tau(x'|x, a)$

(S4) the expected reward given state and action is time-stationary, denoted by $r(s, a)$.

We use this simple MDP to capture the delayed effect on the future rewards of sending the treatment. Note that in this model, the action only impacts the future rewards through the dosage since

the context are assumed independent of the actions and this allows us to form an estimate of delayed effect of treatment based on the current dosage. We assume the context and availability are both i.i.d. across time to reduce the variance of estimating the delayed effect as we do not need to estimate the transition model for the context and availability.

We first discuss how each component in the simple MDP are constructed. Given the history up to the end of day d, H_d, we set (1) the average prior availability is $p_{\text{avail}} = \frac{1}{5d}\sum_{k,l=1}^{d,5} I_{l,k}$, (2) the empirical distribution on $\{Z_{l,k}\}$ is $F(\cdot) = \frac{1}{5d}\sum_{k,l=1}^{d,5}\delta_{Z_{l,k}}(\cdot)$ where $\delta_z(\cdot)$ is the Dirac measure, and (3) the reward function at available decision times is $r(s,a) = g(s)^\top\hat{\alpha}_0 + af(s)^\top\hat{\beta}$ where $\hat{\alpha}_0$, $\hat{\beta}$ are the posterior means based on the model 6.3. The mean reward at unavailable decision times has the same form but with posterior means from a similar linear Bayesian regression using the unavailable time points in H_d. To complete the description of the MDP, we need to specify the transition model, $\tau(x'|x,a)$ for the dosage variable $\{X_t\}$. Recall that the dosage variable is defined at the beginning of section 6.5. Let p_{sed} be the probability of delivering any anti-sedentary suggestions between decision times given no activity suggestion was sent at the previous decision time. We set $p_{\text{sed}} = 0.3$ based on the planned scheduling of anti-sedentary suggestions (an average of 1.5 anti-sedentary suggestion uniformly distributed in a 12-hour time window during the day). Then $\tau(x'|x,a)$ is given by $\tau(x'|x,1) = \mathbb{1}_{\{x'=\lambda x+1\}},\ \tau(x'|x,0) = p_{\text{sed}}\mathbb{1}_{\{x'=\lambda x+1\}} + (1-p_{\text{sed}})\mathbb{1}_{\{x'=\lambda x\}}$. Recall from section 6.5 that $\lambda = 0.9$.

We formulate the proxy of delayed effect based on the above constructed MDP as follows. Consider an arbitrary policy π that chooses the action $\pi(s)$ at the state $S = (Z, I, X)$ if available (i.e., $I = 1$) and chooses action 0 otherwise. Recall the state-action value function for policy π under discount rate γ:

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[R_1 + \gamma R_2 + \gamma^2 R_3 + \dots \mid S_1 = s, A_1 = a\right]$$

where the subscript π means the actions $(A_2, A_3, \dots)$ are selected according to the policy π. We will break Q^π into two parts: $Q^\pi(s,a) = r(s,a) + \gamma H^\pi(x,a)$ where $r(s,a)$ is the expected reward in (S4) and $H^\pi(x,a) = \mathbb{E}_\pi[R_2+\gamma R_3+\gamma^2 R_4+\dots \mid S_1 = s, A_1 = a]$ is the sum of future discounted rewards (future value for short) which excludes the first, immediate reward (R_1) and is only a function of x under the working assumptions (S1) and (S2). Note that the difference $H^\pi(x,1) - H^\pi(x,0)$ measures the impact of sending treatment at dosage x onto the future rewards when the future actions are selected by the policy π. To select the policy, we choose the one that maximizes the future value while restricting to the policy that only depends on the dosage and availability. Specifically, let $H^*(x,a) = \max\{H^\pi(x,a) : \pi : \mathcal{X}\times\{0,1\} \to \mathcal{A},\ \pi(x,0) = 0, \forall x \in \mathcal{X}\}$. It can be shown that H^* is given by $H^*(x,a) = \sum_{x'}\tau(x'|x,a)(p_{\text{avail}}U_1^*(x') + (1-p_{\text{avail}})U_0^*(x'))$, where

U_0^* and U_1^* solves the following equation for all $x \in \mathcal{X}$:

$$U_1(x) = \max_a \left\{ r_1(x,a) + \gamma \sum_{x'} \tau(x'|x,a)(p_{\text{avail}} U_1(x') + (1 - p_{\text{avail}}) U_0(x')) \right\}$$
$$U_0(x) = r_0(x) + \gamma \sum_{x'} \tau(x'|x,0)(p_{\text{avail}} U_1(x') + (1 - p_{\text{avail}} U_0(x')),$$

where r_0 and $r_1(x,a)$ are the marginal reward function (e.g., marginal in the sense that it only depends on the dosage variable) given by

$$r_0(x) = \int r((z,0,x),0) dF(z),\ r_1(x,a) = \int r((z,1,x),a) dF(z)$$

Finally, the proxy for the delayed effect is calculated by

$$\eta_{d+1}(x) = (1-\gamma)[H_{d+1}(x,0) - H_{d+1}(x,1)] \tag{6.5}$$

where $H_{d+1} = (1-w)H_{\text{init}} + wH^*$ is the weighted average between the estimate H^* and the initial function H_{init} calculated based on only data from HeartSteps V1. The selection of the discount rate γ and the weight w will be discussed in section 6.5. This delayed effect is the average difference between the average future value between sending nothing vs. an activity suggestion. The use of $(1-\gamma)$ is to standardize the sum of the discounted rewards e.g., $H_{d+1}(x,a)$ is, on average, the sum of undiscounted $(1-\gamma)^{-1}$ number of rewards.

Choosing Inputs

We review the inputs required by HeartSteps V2 RL algorithm and discuss how each is selected based on the data from HeartSteps V1. The list of required inputs can be found in Figure 3.

First, the scientific team decided $\epsilon_0 = 0.2$ and $\epsilon_1 = 0.1$ in the probability clipping to ensure enough exploration, e.g. forcing the RL algorithm continuously explore without locking into a deterministic policy. As mentioned in section 6.5, we define the dosage in the form of $X_{t+1} = \lambda X_t + \mathbb{1}_{E_{t+1}}$ (recall this variable is used to form the proxy for the delayed effect (6.2). Generalized Estimating Equations' (GEE) analysis [94] was conducted using HeartStep V1 data for a variety values of λ. When λ is relatively large the dosage significantly impacts the effect of the message on the subsequent 30 minute step count. The scientific team selected $\lambda = 0.9$.

In the nightly posterior updates of treatment effect estimates, the working model (6.3) requires the features vectors, $f(s)$ and $g(s)$ in (6.1) and (6.4) , the variance of the noise, σ^2 and the prior

distribution, $\boldsymbol{N}(\mu_{\alpha_0}, \Sigma_{\alpha_0})$ and $\boldsymbol{N}(\mu_\beta, \Sigma_\beta)$. We discuss how to choose them using HeartSteps V1 data in the followings.

First, most of the features are selected based on the GEE results using HeartSteps V1 data. For example, we found that although the 30-minute step count prior to the decision is highly predictive of the rewards (e.g., 30 minute step count after the decision), it is not significant in terms of predicting the treatment effect. Therefore, the prior 30-minute step count is included in the baseline features $g(s)$, but not in the feature vector $f(s)$ for treatment effect. A measure of how participant engages with the mobile app is planned to include in both $g(s)$ and $f(s)$. This variable was not collected in HeartSteps V1. The scientific team believes this variable likely interacts with the treatment and thus decide to include into the features. The features in the feature vector $f(s)$ (6.1) are dosage, location and the variation level of step count 60 minutes around the current time slot in past 7 days. These features along with the prior 30-minute step count, yesterday's total step count and current temperature are included in the baseline feature vector, $g(s)$.

Second, about the variance of the noise σ^2. Although σ^2 can be learned on the fly, e.g., the residual variance by fitting the model using the data collected from the participant, to ensure the stability of the algorithm (e.g., the step count can be highly noisy), we set the variance parameter using the data from HeartSteps V1, that is, σ^2 is not updated during the study.

Third, the prior is constructed based on the analysis result in HeartSteps V1. Specifically, we first conduct Generalized Estimating Equations' (GEE) regression analyses [94], using all participants' data in HeartStep V1 and assess the significance of each feature. To form the prior variance, on each participant we fit a separate GEE linear regression model and calculated the standard deviations of the point estimates across the 37 participant models. We formed the prior mean and prior standard deviation as follows: (1) For the features that are significant in the GEE analysis using all participants' data, we set the prior mean to be the point estimate from this analysis; we set the prior standard deviation to the standard deviation across participant models from the participant specific GEE analyses. (2) For the features that are not significant, we set the corresponding prior mean to be zero and shrink the standard deviation by half. $\Sigma_{\alpha_0}, \Sigma_\beta$ are diagonal matrices with the above prior variances on the diagonals. The same procedure is applied to form the prior mean and variance for the reward model at the unavailable times, used in the proxy value updates. The rationale of setting the mean to zero and shrinking the standard deviation for the non-significant features is to ensure the stability of the algorithm: unless during the HeartSteps V2 study there is strong evidence or signal detected from the participant, these features only have minimal impact on the selection of actions.

The estimates of proxy delayed effect, e.g., η_{d+1}, requires the initial proxy value estimates

H_{init}. To calculate H_{init} we use the same procedure as described in the section 6.5 to calculate H^*, except that the empirical probability of being available, the empirical distribution of contexts and the reward function are constructed only using HeartSteps V1 data.

Three remaining parameters in the HeartSteps V2 RL algorithm need to be specified: the discount rate γ, the updating weight parameter w (both part of the proxy MDP in section 6.5), and the weight parameter ξ in the action selection (6.2). For simplicity, we call them as "tuning parameters" in the rest of the chapter and we will discuss how to select these parameters in the next section.

Selecting Tuning Parameters The tuning parameters, e.g., (w, ξ, γ), are chosen based on a simulation-based procedure: we build a simulation environment (e.g., the data generative model) based on HeartSteps V1 data, apply the algorithm as shown in Figure 3 and then choose the set of parameters that maximizes the total simulated rewards.

We create the simulation environment as follows. Recall that HeartSteps V1 is a 42-day study. For each participant's data in HeartSteps V1, we first fit a person-specific regression model with the feature vector $g(s)$ and $g(S)$and obtain the residuals $\{\tilde{\epsilon}_t\}_{t=1}^{210}$ for all decision times (e.g. $42 \times 5 = 210$). We extract the 42-day sequence of the context, availability, residual, $\{Z_t, I_t, \tilde{\epsilon}_t\}_{t=1}^{210}$. To mimic the HeartSteps V2 study, we extend the sequence to be of 90 days by concatenating (90-42) randomly sampled days' data from the 42-day data in HeartSteps V1. Denote the extended sequence by $\{Z_t, I_t, \tilde{\epsilon}_t\}_{t=1}^{450}$. Note that the sequence is created for all participants (e.g., we have 37 sequences of context-availability-residuals; recall that there are 37 participants in HeartSteps V1 study). In the simulation, the actions are selected by the RL algorithm and the dosage is generated according to the definition in section 6.5 by randomly distributing the anti-sedentary messages. Now we specify how to generate the rewards. To this end, we use all participants' data and fit a population regression model with the linear feature vectors $f(s)$ and $g(s)$. Denote the corresponding estimates by $\tilde{\alpha}$ and $\tilde{\beta}$. Given the current state $S_t = \{Z_t, I_t, X_t\}$ and action A_t , the reward is generated by $R_t = g(S_t)^\top \tilde{\alpha} + A_t f(S_t)^\top \tilde{\beta} + \tilde{\epsilon}_t$. This gives essentially 37 generative models, indexed by the participant.

For each given set of tuning parameters (w, ξ, γ), we run the RL algorithm, as displayed in Figure 3, 128 times for each participant's generative model (i.e., a total of 37×128 times) with the rest of the inputs specified according to the procedure in the beginning of section 6.5. The average of the total rewards that are collected in each run is calculated. We use a grid search, e.g. $w \in \{0, 0.25, 0.5, 0.75, 1\}$, $\gamma \in \{0, 0.5, 0.8, 0.9, 0.95\}$, and $\xi \in \{0, 0.01, 0.05, 0.1, 0.15, 0.2\}$, to select the ones that maximize the average total rewards.

6.6 Simulation Study

We conduct a simulation study to inform the expected performance of the planned HeartSteps V2 RL algorithm and compare with the standard linear Thompson Sampling bandit algorithm, by using HeartSteps V1 data. To this end, we split the data into three folds; each fold contains about 12 participants' data. In each of the three iterations, two of the three folds are used as "training set" and the remaining one fold is used as "testing set". We use the training set to select the inputs (e.g., the prior distribution) and the tuning parameters, (γ, w, ξ) in the algorithm, as described in section 6.5. We then use the testing set to create a "testing" simulation environment (this is to distinguish with the simulation environment for selecting the tuning parameters based on the training set) and apply the algorithm with the selected inputs and tunning parameters to select the action. The testing environment is created as described in section 6.5 but using the testing set. For each participant in the testing set, the RL algorithm is re-run 128 times and the average of total collected rewards is calculated. As a comparison, we also implement a standard Thompson Sampling (TS) bandit algorithm with the same probability constraint (e.g., restricting the randomization probabilities of selecting action when available to be within $[0.1, 0.8]$). The inputs to TS bandit, e.g., the prior distribution and the variance parameter are selected by the same procedure described in section 6.5 by using the training dataset.

The same procedure is repeated three times, that is, in each iteration, two folds of data is used to choose the inputs for both HeartSteps RL algorithm and the TS bandit algorithm, and the remaining one fold is used to test the performance. Each participant is assigned to the testing set once and the difference of average total rewards (averaged over 128 runs) is calculated. The simulation results are shown in Figure 6.1. The average improvement of the proposed algorithm over TS bandit, in terms of total rewards over 90 days, is about 21.28 with standard devision 6.85. Note that in the generative models, the immediate treatment effects roughly ranges from 0 to 0.5 (depending on the state). The average improvement of 21.28 is significant.

6.7 Conclusion and Future Work

In this chapter, we developed a Reinforcement Learning algorithm for use in HeartSteps V2. Preliminary validation of the algorithm demonstrates good performance over a standard Thompson Sampling bandit algorithm in synthetic experiments. After HeartSteps V2 is completed, the data will be used to further assess the performance and utility of the algorithm.

We foresee some opportunities for future work. First, our proposed algorithm learns the treat-

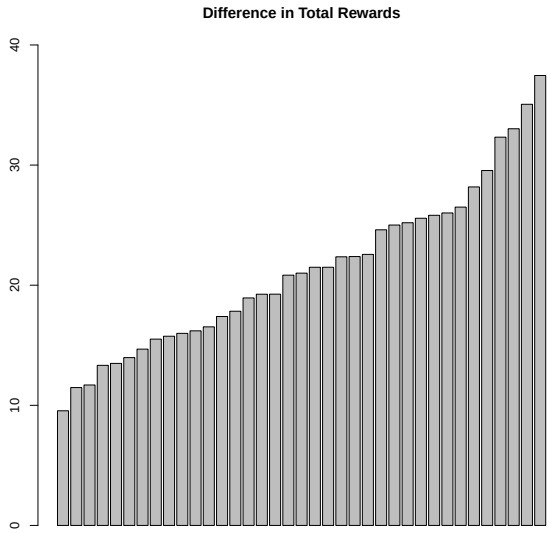

Figure 6.1: Three-fold cross validation result of HeartSteps V2 RL algorithm and Thompson Sampling bandit algorithm. The y-axis is the difference of average total rewards achieved by HeartSteps V2 RL and Thompson Sampling bandit algorithm.

ment policy separately for each participant (e.g. fully personalized). If the participants in the study are similar enough, pooling information from other participants (either currently still in the study or already having finished the study) can speed learning and achieve better performance, especially for those entering the study later. Second, the current algorithm takes into account the delayed effect of treatment by using a pre-defined "dosage variable" capturing the burden. It would be interesting to develop a version in which more sophisticated measures of burden as well as engagement, for example via a latent state model, is used to approximate the delayed effect. Finally, it would be also interesting to investigate how to incorporate the variance in the estimation of the delayed effect in the action selection and investigate theoretically how to best trade-off the exploration and exploitation.

BIBLIOGRAPHY

[1] Liao, P., Klasjna, P., Tewari, A., and Murphy, S., "Micro-Randomized Trials in mHealth," *Statistics in Medicine*, Vol. 35, No. 12, 2016, pp. 1944–71.

[2] Dempsey, W., Liao, P., Kumar, S., and Murphy, S. A., "The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments," *arXiv preprint arXiv:1711.03587*, 2017.

[3] Liao, P., Dempsey, W., Sarker, H., Hossain, S. M., al'Absi, M., Klasnja, P., and Murphy, S., "Just-in-Time but Not Too Much: Determining Treatment Timing in Mobile Health," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, Vol. 2, No. 4, 2018, pp. 179.

[4] Lewis, M. A., Uhrig, J. D., Bann, C. M., Harris, J. L., Furberg, R. D., Coomes, C., and Kuhns, L. M., "Tailored text messaging intervention for HIV adherence: a proof-of-concept study," *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, Vol. 32, No. 3, March 2013, pp. 248—253.

[5] Kaplan, R. M. and Stone, A. A., "Bringing the Laboratory and Clinic to the Community: Mobile Technologies for Health Promotion and Disease Prevention," *Annual Review of Psychology*, Vol. 64, No. 1, 2013, pp. 471–498, PMID: 22994919.

[6] King, A. C., Castro, C. M., Buman, M. P., Hekler, E. B., Urizar, Guido G., J., and Ahn, D. K., "Behavioral Impacts of Sequentially versus Simultaneously Delivered Dietary Plus Physical Activity Interventions: the CALM Trial," *Annals of Behavioral Medicine*, Vol. 46, No. 2, 2013, pp. 157–168.

[7] Marsch, L. A., "Leveraging Technology to Enhance Addiction Treatment and Recovery," *Journal of Addictive Diseases*, Vol. 31, No. 3, 2012, pp. 313–318, PMID: 22873192.

[8] Boyer, E., Fletcher, R., Fay, R., Smelson, D., Ziedonis, D., and Picard, R., "Preliminary Efforts Directed Toward the Detection of Craving of Illicit Substances: The iHeal Project," *Journal of Medical Toxicology*, Vol. 8, No. 1, 2012, pp. 5–9.

[9] Alessi, S. M. and Petry, N. M., "A randomized study of cellphone technology to reinforce alcohol abstinence in the natural environment," *Addiction*, Vol. 108, No. 5, 2013, pp. 900–909.

[10] A. Cucciare, M., R. Weingardt, K., J. Greene, C., and Hoffman, J., "Current Trends in Using Internet and Mobile Technology to Support the Treatment of Substance Use Disorders," *Current Drug Abuse Reviews*, Vol. 5, No. 3, 2012, pp. 172–177.

[11] Gustafson, D., FM, M., M, C., and et al, "A smartphone application to support recovery from alcoholism: A randomized clinical trial," *JAMA Psychiatry*, Vol. 71, No. 5, 2014, pp. 566–572.

[12] Quanbeck, A., Gustafson, D., Marsch, L., McTavish, F., Brown, R., Mares, M.-L., Johnson, R., Glass, J., Atwood, A., and McDowell, H., "Integrating addiction treatment into primary care using mobile health technology: protocol for an implementation research study," *Implementation Science*, Vol. 9, No. 1, 2014, pp. 65.

[13] Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., Patel, V., and Haines, A., "The Effectiveness of Mobile-Health Technology-Based Health Behaviour Change or Disease Management Interventions for Health Care Consumers: A Systematic Review," *PLoS Med*, Vol. 10, No. 1, 01 2013, pp. e1001362.

[14] Nilsen, W., Kumar, S., Shar, A., Varoquiers, C., Wiley, T., Riley, W. T., Pavel, M., and Atienza, A. A., "Advancing the Science of mHealth," *Journal of Health Communication*, Vol. 17, No. sup1, 2012, pp. 5–10.

[15] Muessig, E. K., Pike, C. E., LeGrand, S., and Hightow-Weidman, B. L., "Mobile Phone Applications for the Care and Prevention of HIV and Other Sexually Transmitted Diseases: A Review," *J Med Internet Res*, Vol. 15, No. 1, Jan 2013, pp. e1.

[16] Spruijt-Metz, D. and Nilsen, W., "Dynamic Models of Behavior for Just-in-Time Adaptive Interventions," *Pervasive Computing, IEEE*, Vol. 13, No. 3, July 2014, pp. 13–17.

[17] Kumar, S., Nilsen, W., Pavel, M., and Srivastava, M., "Mobile Health: Revolutionizing Healthcare Through Transdisciplinary Research," *Computer*, Vol. 46, No. 1, 2013, pp. 28–35.

[18] Box, G. E., P.Hunter, J. S., and Hunter, W. G., *Statistics for experimenters : an introduction to design, data analysis, and model building*, Wiley series in probability and mathematical statistics, 1978.

[19] Chakraborty, B., Collins, L. M., Strecher, V. J., and Murphy, S. A., "Developing multicomponent interventions using fractional factorial designs," *Statistics in Medicine*, Vol. 28, No. 21, 2009, pp. 2687–2708.

[20] Rubin, D. B., "Bayesian Inference for Causal Effects: The Role of Randomization," *Ann. Statist.*, Vol. 6, No. 1, 01 1978, pp. 34–58.

[21] Robins, J., "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*, Vol. 7, No. 9–12, 1986, pp. 1393 – 1512.

[22] Robins, J., "Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"," *Computers and Mathematics with Applications*, Vol. 14, No. 9–12, 1987, pp. 923 – 945.

[23] Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F., "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer," *Journal of the American Statistical Association*, Vol. 107, No. 498, 2012, pp. 493–508.

[24] Robins, J. M., "Optimal Structural Nested Models for Optimal Sequential Decisions," *Proceedings of the Second Seattle Symposium on Biostatistics*, edited by D. Y. Lin and P. Heagerty, Springer, New York, 2004, pp. 189–326.

[25] Liang, K.-Y. and Zeger, S. L., "Longitudinal data analysis using generalized linear models," *Biometrika*, Vol. 73, No. 1, 1986, pp. 13–22.

[26] Tu, X. M., Kowalski, J., Zhang, J., Lynch, K. G., and Crits-Christoph, P., "Power analyses for longitudinal trials and other clustered designs," *Statistics in Medicine*, Vol. 23, No. 18, 2004, pp. 2799–2815.

[27] Mancl, L. A. and DeRouen, T. A., "A Covariance Estimator for GEE with Improved Small-Sample Properties," *Biometrics*, Vol. 57, No. 1, 2001, pp. 126–134.

[28] Li, P. and Redden, D. T., "Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes," *Statistics in Medicine*, Vol. 34, No. 2, 2015, pp. 281–296.

[29] Hotelling, H., "The Generalization of Student's Ratio," *Ann. Math. Statist.*, Vol. 2, No. 3, 08 1931, pp. 360–378.

[30] Cohen, J., *Statistical Power Analysis for the Behavioral Sciences(2nd)*, Routledge, 2nd ed., July 1 1988.

[31] Dallery, J. and Raiff, B., "Optimizing behavioral health interventions with single-case designs: from development to dissemination," *Translational Behavioral Medicine*, Vol. 4, No. 3, 2014, pp. 290–303.

[32] Shadish, W. and Sullivan, K., "Characteristics of single-case designs used to assess intervention effects in 2008," *Behavior Research Methods*, Vol. 43, No. 4, 2011, pp. 971–980.

[33] Dallery, J., Cassidy, R., and Raiff, B., "Single-Case Experimental Designs to Evaluate Novel Technology-Based Health Interventions," *Journal of Medical Internet Research*, Vol. 15, 2013, pp. e22.

[34] Sarker, H., Tyburski, M., Rahman, M., Hovsepian, K., Sharmin, M., Epstein, D., Preston, K., Furr-Holden, C., Milam, A., Nahum-Shani, I., al'Absi, M., and Kumar, S., "Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, ACM, Santa Clara, California, USA, 2016, pp. 4489–4501.

[35] Sarker, H., Hovsepian, K., Chatterjee, S., Nahum-Shani, I., Murphy, S., Spring, B., Ertin, E., al'Absi, M., Nakajima, M., and Kumar, S., "From Markers to Interventions: The Case of Just-in-Time Stress Intervention," *Mobile Health Sensors, Analytic Methods, and Applications*, edited by J. Regh, S. Murphy, and S. Kumar, Springer International Publishing, 2017.

[36] Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A., "Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support," *Annals of Behavioral Medicine*, 2016, pp. 1–17.

[37] Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., and Kumar, S., "cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, ACM, New York, NY, USA, 2015, pp. 493–504.

[38] Hossain, S. M., Hnat, T., Saleheen, N., Nasrin, N. J., Noor, J., Ho, B.-J., Condie, T., Srivastava, M., and Kumar, S., "mCerebrum: An mHealth Software Platform for Development and Validation of Digital Biomarkers and Interventions," *The ACM Conference on Embedded Networked Sensor Systems (SenSys)*, ACM, 2017.

[39] Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., and Shah, S., "AutoSense: Unobtrusively Wearable Sensor Suite for Inferring the Onset, Causality, and Consequences of Stress in the Field," *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, New York, NY, USA, 2011, pp. 274–287.

[40] Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S., "Assessing Time-Varying Causal Effect Moderation in Mobile Health," To appear in the *Journal of the American Statistical Association*.

[41] Muhammad, G., Alsulaiman, M., Amin, S. U., Ghoneim, A., and Alhamid, M. F., "A Facial-Expression Monitoring System for Improved Healthcare in Smart Cities," *IEEE Access*, Vol. 5, 2017, pp. 10871–10881.

[42] Zhang, M. W. B., Ward, J., Ying, J. J. B., Pan, F., and Ho, R. C. M., "The alcohol tracker application: an initial evaluation of user preferences," *BMJ Innovations*, Vol. 2, No. 1, 2016, pp. 8–13.

[43] Dulin, P. L., Gonzalez, V. M., and Campbell, K., "Results of a Pilot Test of a Self-Administered Smartphone-Based Treatment System for Alcohol Use Disorders: Usability and Early Outcomes," *Substance Abuse*, Vol. 35, No. 2, 2014, pp. 168–175.

[44] Pielot, M., Cardoso, B., Katevas, K., Serrà, J., Matic, A., and Oliver, N., "Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 1, No. 3, Sept. 2017, pp. 91:1–91:25.

[45] Klasnja, P., Harrison, B. L., LeGrand, L., LaMarca, A., Froehlich, J., and Hudson, S., "Using Wearable Sensors and Real Time Inference to Understand Human Recall of Routine Activities," *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, 2008, pp. 154–163.

[46] Dimitrijević, M., Faganel, J., Gregorić, M., Nathan, P., and Trontelj, J., "Habituation: effects of regular and stochastic stimulation," *Journal of Neurology, Neurosurgery & Psychiatry*, Vol. 35, No. 2, 1972, pp. 234–242.

[47] Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A., "Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps," *Annals of Behavioral Medicine*, 2018.

[48] Klasnja, P., Hekler, E., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S., "Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions." *Health Psychology*, Vol. 34, No. S, 2015, pp. 1220.

[49] Dempsey, W., Liao, P., Nahun-Shani, P. K. I., and Murphy, S., "Randomised trials for the Fitbit generation," *Significance*, Vol. 12, No. 6, 2015, pp. 20–23.

[50] Dempsey, W., Liao, P., and Murphy, S., "Sample size calculations for stratified micro-randomised trials in mHealth," Submitted.

[51] Shiffman, S., Stone, A. A., and Hufford, M. R., "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, Vol. 4, 2008, pp. 1–32.

[52] Scott, C. K., Dennis, M. L., Gustafson, D., and Johnson, K., "A pilot study of the feasibility and potential effectiveness of using smartphones to provide recovery support," *Drug & Alcohol Dependence*, Vol. 171, 2017, pp. e185.

[53] Scott, C., Dennis, M., and Gustafson, D., "Using smartphones to decrease substance use via self-monitoring and recovery support: study protocol for a randomized control trial." *Trials*, Vol. 18, No. 1, 2017, pp. 374.

[54] Dennis, M., Scott, C. K., Funk, R. R., and Nicholson, L., "A pilot study to examine the feasibility and potential effectiveness of using smartphones to provide recovery support for adolescents," *Substance abuse*, Vol. 36, No. 4, 2015, pp. 486–492.

[55] Rathbun, S., Song, X., Neustfiter, B., and Shiffman, S., "Survival analysis with time-varying covariates measured at random times by design." *J R Stat Soc Ser C Appl. Stat.*, Vol. 62, No. 3, 2012, pp. 419–434.

[56] Stone, A., Shiffman, S., Atienza, A., and Nebeling, L., *The science of real-time data capture: Self-reports in health research*, Oxford University Press, 2007.

[57] Wen, C., Schneider, S., Stone, A., and Spruijt-Metz, D., "Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis," *J Med Internet Res*, Vol. 19, No. 4, 2017, pp. e132.

[58] De Gooijer, J. G. and Hyndman, R. J., "25 years of time series forecasting," *International journal of forecasting*, Vol. 22, No. 3, 2006, pp. 443–473.

[59] Wikipedia, "KullbackLeibler divergence — Wikipedia, The Free Encyclopedia," http://en.wikipedia.org/wiki/KullbackLeibler_divergence, 2018, [Online; accessed 26-July-2018].

[60] Saleheen, N., Ali, A., Hossain, S., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., and Kumar, S., "puffMarker: A Multi-sensor Approach for Pinpointing the Timing of First Lapse in Smoking Cessation," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, ACM, New York, NY, USA, 2015, pp. 999–1010.

[61] Bradtke, S. J., Barto, A. G., and Kaelbling, P., "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, 1996, pp. 22–33.

[62] Nedić, A. and Bertsekas, D. P., "Least squares policy evaluation algorithms with linear function approximation," *Discrete Event Dynamic Systems*, Vol. 13, No. 1-2, 2003, pp. 79–110.

[63] Ueno, T., Kawanabe, M., Mori, T., Maeda, S.-i., and Ishii, S., "A semiparametric statistical approach to model-free policy evaluation," *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 1072–1079.

[64] Lazaric, A., Ghavamzadeh, M., and Munos, R., "Finite-sample analysis of least-squares policy iteration," *Journal of Machine Learning Research*, Vol. 13, No. Oct, 2012, pp. 3041–3074.

[65] Tagorti, M. and Scherrer, B., "On the Rate of Convergence and Error Bounds for LSTD (λ)," *International Conference on Machine Learning*, 2015, pp. 1521–1529.

[66] Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S., "Regularized policy iteration with nonparametric function spaces," *The Journal of Machine Learning Research*, Vol. 17, No. 1, 2016, pp. 4809–4874.

[67] Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R., "Estimating dynamic treatment regimes in mobile health using V-learning," *Journal of the American Statistical Association*, , No. just-accepted, 2019, pp. 1–39.

[68] Puterman, M. L., "Markov Decision Processes: Discrete Stochastic Dynamic Programming," 1994.

[69] Hernández-Lerma, O. and Lasserre, J. B., *Further topics on discrete-time Markov control processes*, Vol. 42, Springer, 1999.

[70] Ortner, R. and Ryabko, D., "Online regret bounds for undiscounted continuous reinforcement learning," *Advances in Neural Information Processing Systems*, 2012, pp. 1763–1771.

[71] Van de Geer, S., *Empirical Processes in M-estimation*, Vol. 6, Cambridge university press, 2000.

[72] Zhao, T., Cheng, G., and Liu, H., "A partially linear framework for massive heterogeneous data," *Annals of statistics*, Vol. 44, No. 4, 2016, pp. 1400.

[73] Steinwart, I. and Christmann, A., *Support vector machines*, Springer Science & Business Media, 2008.

[74] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H., *A distribution-free theory of nonparametric regression*, Springer Science & Business Media, 2006.

[75] Munos, R., "Performance bounds in l_p-norm for approximate value iteration," *SIAM journal on control and optimization*, Vol. 46, No. 2, 2007, pp. 541–561.

[76] Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I., "Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system," *Journal of medical Internet research*, Vol. 19, No. 10, 2017.

[77] Paredes, P., Gilad-Bachrach, R., Czerwinski, M., Roseway, A., Rowan, K., and Hernandez, J., "PopTherapy: coping with stress through pop-culture," *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 109–117.

[78] Forman, E. M., Kerrigan, S. G., Butryn, M. L., Juarascio, A. S., Manasse, S. M., Ontañón, S., Dallal, D. H., Crochiere, R. J., and Moskow, D., "Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss?" *Journal of behavioral medicine*, 2018, pp. 1–15.

[79] Rabbi, M., Aung, M. H., Zhang, M., and Choudhury, T., "MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2015, pp. 707–718.

[80] Zhou, M., Mintz, Y., Fukuoka, Y., Goldberg, K., Flowers, E., Kaminsky, P., Castillejo, A., and Aswani, A., "Personalizing Mobile Fitness Apps using Reinforcement Learning," *Companion Proceedings of the 23rd International on Intelligent User Interfaces: 2nd Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE)*, 2018.

[81] Ghosh, A., Chowdhury, S. R., and Gopalan, A., "Misspecified Linear Bandits." *AAAI*, 2017, pp. 3761–3767.

[82] Dimakopoulou, M., Athey, S., and Imbens, G., "Estimation Considerations in Contextual Bandits," *arXiv preprint arXiv:1711.07077*, 2017.

[83] Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y., "Non-Stationary Bandits with Habituation and Recovery Dynamics," *arXiv preprint arXiv:1707.08423*, 2017.

[84] Thomas, P. and Brunskill, E., "Data-efficient off-policy policy evaluation for reinforcement learning," *International Conference on Machine Learning*, 2016, pp. 2139–2148.

[85] Jiang, N. and Li, L., "Doubly robust off-policy value evaluation for reinforcement learning," *arXiv preprint arXiv:1511.03722*, 2015.

[86] Chapelle, O. and Li, L., "An empirical evaluation of thompson sampling," *Advances in neural information processing systems*, 2011, pp. 2249–2257.

[87] Osband, I. and Van Roy, B., "Why is Posterior Sampling Better than Optimism for Reinforcement Learning?" *arXiv preprint arXiv:1607.00215*, 2016.

[88] Osband, I. and Van Roy, B., "On optimistic versus randomized exploration in reinforcement learning," *arXiv preprint arXiv:1706.04241*, 2017.

[89] Greenewald, K., Tewari, A., Murphy, S., and Klasnja, P., "Action centered contextual bandits," *Advances in neural information processing systems*, 2017, pp. 5977–5985.

[90] Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V., "Semiparametric Contextual Bandits," *arXiv preprint arXiv:1803.04204*, 2018.

[91] Osband, I., Russo, D., and Van Roy, B., "(More) efficient reinforcement learning via posterior sampling," *Advances in Neural Information Processing Systems*, 2013, pp. 3003–3011.

[92] Fonteneau, R., Korda, N., and Munos, R., "An optimistic posterior sampling strategy for Bayesian reinforcement learning," *NIPS 2013 Workshop on Bayesian Optimization (BayesOpt2013)*, 2013.

[93] Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R., "Learning unknown Markov decision processes: A Thompson sampling approach," *Advances in Neural Information Processing Systems*, 2017, pp. 1333–1342.

[94] Liang, K. and Zeger, S., "Longitudinal data analysis using generalized linear models," *Biometrika*, Vol. 73, No. 1, 1986, pp. 13–22.