EXTENDING DYNAMIC TREATMENT REGIMES TO INCORPORATE LONGITUDINAL DATA OBSERVED BETWEEN DECISION TIMES

Mengbing Li

Senior Honors Thesis

Department of Biostatistics

University of North Carolina at Chapel Hill

2017

Approved by:

Dr. Michael Kosorok, Thesis Advisor,

Dr. Jane Monaco, Reader,

Dr. Gary Koch, Reader,

EXTENDING DYNAMIC TREATMENT REGIMES TO INCORPORATE LONGITUDINAL DATA OBSERVED BETWEEN DECISION TIMES

Mengbing Li

Senior Honors Thesis

Department of Biostatistics

University of North Carolina at Chapel Hill

2017

Approved by:

Dr. Michael Kosorok, Thesis Advisor,

Dr. Jane Monaco, Reader,

Dr. Gary Koch, Reader,

ABSTRACT

MENGBING LI: Extending Dynamic Treatment Regimes to Incorporate Longitudinal Data Observed Between Decision Times.
(Under the direction of Dr. Michael Kosorok, Thesis Advisor)

Personalized medicine refers to the medical scheme that tailors treatment to individuals based on individual characteristics, predicted risks, and expected outcomes. Two important components of personalized medicine involve the estimation of individualized treatment rules (ITRs) and the design of adaptive clinical trials. Dynamic treatment regimes (DTRs) are sequential treatment rules for individual patients that are adaptive over their disease progresses. Much research on estimation of the optimal DTRs has been carried out in the recent decade, and machine learning methods have been employed in the estimation. It should be noted that when estimating the optimal DTRs, we usually face the issue of sparsity in asynchronously collected data, which standard statistical methods for longitudinal data may not be applicable. In this thesis, we first review existing two major machine learning methods, Q-learning and outcome weighted learning, that are applicable to estimating ITRs with longitudinal data. Then we propose a new learning method that deal with asynchronous sparse longitudinal data when the treatment option is binary. This method uses a counting process to generate new features, and then utilizes a Q-learning-like approach to estimate parameters in the decision function. We also discuss advantages and limitations of the proposed method, as well as possible directions of future research.

ACKNOWLEDGMENTS

I would first like to thank my parents who have offered me emotional and financial support throughout college. Without them I may not have found myself at UNC, nor had the courage to pursue whatever subjects I enjoy or to engage in this task.

Importantly, I would like to thank my thesis advisor Professor Michael Kosorok, for the guidance and advice throughout the process of writing the thesis. I am very grateful for the precious time and extraordinary patience he offered me when helping me learn the very basics of machine learning and its applications in clinical trials, guiding me to read and understand research papers, helping me narrow down my thesis topic, and providing insights and editorial support for my writing. Thank you Dr. Kosorok, for showing an interesting brand new research area to me and giving me a good taste of research.

Next, I would like to thank both Professor Gary Koch and Professor Jane Monaco. Since my junior year, Dr. Koch has been providing me a position at the Biometrics Consulting Lab, where I had the opportunity to get involved in statistical projects that allowed me to practice the knowledge I learned from classes in real world settings. He has also offered me numerous useful and insightful suggestions on course selection, career planning, and graduate school applications. I have known Dr. Monaco since my sophomore year, and it was she and her passion that initially inspired me to pursue biostatistics. Without her generous help and invaluable advice throughout the past three years, my study at UNC would not have been so smooth and enjoyable.

Lastly, thank you to my friends and roommates for bringing me wonderful life experiences. Thank you to my professors in the STOR department who I have taken classes with and who attract me to the wonderful world of statistics. Finally, a big thank you to UNC-Chapel Hill, of which I am always proud being a Tar Heel.

TABLE OF CONTENTS

LIST OF FIGURES vi

1 INTRODUCTION 1

2 DATA SETTINGS AND NOTATIONS 5

2.1 Standard Settings 5

2.1.1 Individualized Treatment Rule in Standard Single-Stage Settings 5

2.1.2 Dynamic Treatment Regimes in Standard Multi-Stage Settings 5

2.1.3 Observational Setting 8

2.2 Dynamic Treatment Regimes (DTRs) with Additional Longitudinal Data 9

2.2.1 Regularly Spaced Data 9

2.2.2 Irregularly Spaced Data 11

3 REINFORCEMENT LEARNING 13

3.1 Reinforcement Learning and Q-Learning Backgrounds 13

3.2 Estimating the Q-Function 16

3.2.1 Support Vector Regression 16

3.2.2 Extremely Randomized Trees 18

3.3 Discussion 18

4 OUTCOME WEIGHTED LEARNING 19

5 BACKWARD AND SIMULTANEOUS OUTCOME WEIGHTED LEARNING 23

5.1 Backward Outcome Weighted Learning (BOWL) 23

5.2 Simultaneous Outcome Weighted Learning 25

5.3 Discussion 26

6 PROPOSED METHODOLOGY FOR ADDITIONAL LONGITUDINAL DATA .. 27
7 DISCUSSION 31
8 REFERENCES 33

LIST OF FIGURES

2.1 Figure2.1 8
2.2 Figure2.2 10
2.3 Figure2.3 12

1 INTRODUCTION

Many clinical trials are designed to examine drug effects on the patient population as a whole in a single stage. However, this unchanged "one-size-fits-all" scheme can be problematic in clinical practice because of heterogeneity of patient characteristics and differences in patient treatment progression. An ideal optimal treatment regime is expected to overcome such problems and be individualized and adaptive over time. For example, in treating patients with psychiatric disorders, clinicians need to consider individual characteristics which may influence treatment response. Considering the delayed treatment effects and potential reoccurrence of symptoms, clinicians may also want to relieve the waxing and waining of patients following long-term treatments, which significantly increase the risks of severe side effects, psychological and physical stress, as well as economic burden of patients (Murphy *et al.* 2006). Therefore, a tailored adaptive treatment design will contribute to both optimizing treatment effects as well as reducing patient burden.

Dynamic treatment regimes (DTRs), also called adaptive treatment strategies (Murphy 2005), are a general approach to address these concerns. DTRs refer to a sequence of treatments tailored to a set of covariates, including individual patient characteristics, dosage level, treatment response, etc. that may or may not be changing over time. Patient status is reevaluated at each predetermined time point, which breaks up the entire treatment into many stages, and the treatment given to the patient is decided based on all covariate history up to that time. The ultimate goal of DTRs is to find a treatment regime that maximizes the average expected outcome provided that the whole population follow that regime (Kidwell 2015). Under this framework, the one-size-fits-all issue and delayed effects are successfully addressed. The flexibility and maximal benefit have made DTRs increasingly popular in

clinical practice.

However, although DTRs require patient data collected at each decision point, patient status and disease progress are not constantly monitored between stages. Additional data within each stage over time, whether regularly or irregularly collected or not, may aid in finding better treatment regimes for patients. In a recently conducted study, researchers appealed to adaptive mobile health (mHealth) to help smokers quit (McClure *et al.* 2016). In the study, participants randomized to an adaptive interactive program which provided real-time, adaptively tailored advice on top of standard self-help content, showed a higher proportion (76%) of quitting than participants randomized to an non-adaptive program with standard self-help content (67%). Thus mHealth can serve as a promising intervention to provide additional data along with treatment in that smart phones are capable of offering prompt feedback to both smokers and clinicians to make better tailored treatment decisions.

Another example that illustrates the usefulness of collecting additional patient data between stages is treatment for Type I diabetes (T1D). Young adults with T1D often struggle with glycemic control and weight management (Liu *et al.* 2010). Compared to other adults with T1D, those young adults are more likely to experience extra energy loss resulting from glucosuria (Anderbro *et al.* 2010), increased resting energy expenditure (Schober *et al.* 2011), as well as increased level of metabolic activity (Wadden *et al.* 2002) due to their physiological characteristics. In addition, young adults with T1D are exposed to an increased risk of weight gain resulting from intense hunger, which is associated with recurrent hypoglycemia and poor glycemic self-control (Pinhas-Hamiel and Levy-Shraga 2013). Usually, these patients visit physicians on a regular basis and seek feedback and updated treatment options based on their status at the visit. With the above factors in mind, monitoring real-time energy expenditure and weight change between their clinical visits is potentially helpful for physicians to make better decisions on treatments. Thus it can be expected that collecting additional longitudinal data on patient status is a useful approach

to optimizing glycemic control and weight management.

Although the scheme may seem promising, carrying out such a process requires many efforts. On one hand, even though rapidly developing technologies greatly facilitate the data collection processes through, for example, smart phones, electronic wristbands, and other portable devices, there is no existing protocol specifying rules and restrictions of the data collection process. On the other hand, challenges of analyzing the data are significant.

First, finding the optimal DTR usually requires machine learning methods. In (Zhao *et al.* 2009), the authors applied reinforcement learning, specifically Q-learning, to discovering the optimal treatment regime in clinical trials for life-threatening diseases. Q-learning is a model-free temporal difference learning algorithm that deals with infinite-state Markov Decision Processes (MDP). Rather than learning the MDP, Q-learning instead learns the value of each state and the optimal policy directly by only using existing states and available actions in each state. The goal of Q-learning is to optimize the Q-function, which is the expected discounted reward after executing an action at the current state and following the policy in all states afterwards. Support vector regression and extremely randomized trees were used to estimate the Q-function. The method did not rely on precise dynamic mathematical models, and successfully incorporated delayed effects of treatments, drug efficacy, and drug toxicity into improving long-term clinical outcomes.

(Zhao *et al.* 2012) was among the first to use machine learning techniques for classification in estimating optimal treatment rules. The authors proposed outcome weighted learning (OWL) in estimating individualized treatment rules (ITR) with binary options. The method directly finds the optimal ITR that maximizes the clinical outcome using prognostic variables without modeling the conditional means. (Zhao *et al.* 2015) proposed two new nonparametric machine learning methods for estimating the optimal DTR. One is called backward outcome weighted learning (BOWL), which treats estimating the optimal DTR as a sequence of weighted classification problems. It starts from the last stage, estimating

the optimal decision rule in future stages first and then the optimal decision rule in the preceding stages by restricting analysis to patients who followed exactly all future treatment rules. The other is called simultaneous outcome weighted learning (SOWL), which sees estimating the optimal DTR as a single classification problem. It finds the optimal DTR by directly maximizing the expected average reward. All the above methods are useful in a standard DTR design, but may not be directly applicable to the newly proposed scheme where additional longitudinal data are collected between stages.

Second, since we would like to monitor real-time changes in patient status, it is very likely that data are collected at irregularly spaced time points and are distributed sparsely (Cao *et al.* 2015). Given a patient, the sparsity refers to the small number of covariates and response variables that are observed at the same time, leading to asynchronous data and violating assumptions for standard methods for analyzing longitudinal data. Thus special methods such as the one proposed by (Cao *et al.* 2015) should be considered.

The rest of this thesis is organized as follows. In chapter 2, we introduce individualized treatment rule for the single-stage and multiple-stage decision settings. We also introduce the general setting of dynamic treatment regimes with additional data collected between stages. Each setting will be discussed separately for regularly spaced data and irregularly spaced data. In chapters 3 and 4, we present Q-learning and outcome weighted learning respectively, which are existing machine learning methods for estimating the optimal DTR using. Chapter 5 presents two new algorithms based on outcome weighted learning that estimate the optimal DTR sequentially and simultaneously, respectively. In chapter 6, we discuss potential methods to deal with the sparsity in our data. Potential extensions and future are discussed in chapter 7.

2 DATA SETTINGS AND NOTATIONS

2.1 Standard Settings

2.1.1 Individualized Treatment Rule in Standard Single-Stage Settings

In the usual standard setting, we only collect patient data at a single time point. We denote available information of each patient as a tuple (X_n, A_n, Y_n), where $n = 1, \ldots, N$, and each tuple is an independent and identically distributed trajectory of (X, A, Y). Here X is a p-dimensional random vector of covariates. We consider a setting where treatment assignments $A \in \mathcal{A}$ are independent of patient covariates X, where $\mathcal{A}$ is the collection of treatments received. $\mathcal{A}$ can be of any form including binary, discrete, and continuous. Y is the observed clinical outcome, which may also be called the reward depending on the context, and is coded so that larger values correspond to better outcomes. We assume Y is bounded. Let $\mathcal{D}$ be the collection of all possible treatment rules. An individualized treatment rule (ITR) is a map $d : \mathcal{X} \to \mathcal{A}$. An optimal ITR, denoted as d^{opt}, is a rule that maximizes the expected outcome if implemented by the entire population. Thus our goal is to quantify the relationship between (X, A, Y) so that the maximum Y can be achieved. Depending on the type of treatment assignments, we can apply regression methods, such as the generalized linear regression, to establish the desired relationship.

2.1.2 Dynamic Treatment Regimes in Standard Multi-Stage Settings

In a K-stage setting, we collect patient data at K decision time points. We can represent each patient's available information as $(X_{n1}, A_{n1}, X_{n2}, A_{n2}, \ldots, X_{nK}, A_{nK}, Y_n)$ where $n = 1, \ldots, N$ and each tuple is an independent and identically distributed trajectory of $(X_1, A_1, X_2, A_2, \ldots, X_K, A_K, Y)$ sampled at random from a distribution P. Here for each $k = 1, \ldots, K$, $A_k \in \mathcal{A}_k$ is the treatment assignment at the k^{th} stage where $\mathcal{A}_k$ is the

collection of treatment assignments at stage k. X_k is the available patient information after treatment assignment A_{k-1} but prior to the k^{th} stage. Y is the final outcome after all stages of treatments and is coded so that larger values correspond to better outcomes. Let $H_k = (X_1, A_1, \ldots, A_{k-1}, X_k) \in \mathcal{H}_k$ be the history information up to stage k with $H_1 = X_1$, and $\mathcal{D}_k$ be the collection of available treatments at stage k. A dynamic treatment regime (DTR) is a sequence of decision rules $\mathbf{d} = (d_1, \ldots, d_K)$ where each d_k is a map from $\mathcal{H}_k$ to $\mathcal{D}_k$. Our goal is to find the optimal DTR d^{opt} that maximizes the expected average outcome if the rule is implemented by the entire population in the future (Zhao 2015).

We can formalize the process through potential outcomes. We will use lowercase letters a_k to denote the realized treatment at stage k. An overbar will be used to denote events that happened in the past, and an underbar will be used to denote events that will happen in the future. Thus we have $\bar{a}_k = (a_1, a_2, \ldots, a_k)$, and $\underline{a}_k = (a_k, a_{k+1}, \ldots, a_K)$. Note that $\bar{d} = \mathbf{d}$. Let $X^*(\bar{a}_k)$ be a patient's potential covariate status at the start of stage k provided the sequence of treatments $(a_1, a_2, \ldots, a_k)$ was assigned. Let $Y^*(\bar{a}_K)$ be a patient's potential outcome at the end of the study provided the sequence of treatments $(a_1, \ldots, a_K)$ was followed. In the above framework, we can write $h_1 = x_1, a_1 = d_1(x_1), x_2 = X^*(d_1) = X^*(a_1), h_2 = (x_1, a_1, x_2) = (\bar{x}_2, a_1), a_2 = d_2(h_2), x_3 = X^*(\bar{d}_2) = X^*(\bar{a}_2), \ldots, a_{K-1} = d_{K-1}(h_{K-1}), x_K = X^*(\bar{d}_{K-1}) = X^*(\bar{a}_{K-1}), h_K = (\bar{x}_K, \bar{a}_{K-1}), a_K = d_K(h_K)$ and $Y^*(\overline{a_K}) = Y^*(\bar{d}) = Y^*(\mathbf{d})$. Our optimal DTR will be a rule with the property $\mathbb{E}\{Y^*(\mathbf{d})\} \leq \mathbb{E}\{Y^*(d^{opt})\}$ for any $\mathbf{d} \in \{(d_1, \ldots, d_K) \mid d_k \in \mathcal{D}_k\}$.

In order analyze the DTR setup mathematically, we need several assumptions: (1) causal consistency, or stable unit treatment value assumption (SUTVA) (2) sequential ignorability, or no unmeasured confounders, or conditional exchangeability; (3) positivity (Zhao 2015).

1. Causal consistency: we assume that the potential outcome under a sequence of treatments is the same as the observed outcome under this sequence of assigned treatments. Mathematically we can express this assumption as for $\forall k = 1, \ldots, K$, if $\overline{A}_{k-1} = \bar{a}_{k-1}$ then $X_k = X^*(\bar{a}_{k-1})$, and if $\overline{A}_K = \bar{a}_K$ then $Y = Y^*(\bar{a}_K)$.

2. Sequential ignorability: we assume that given the history information of patient covariates and treatment assignments up to stage k, the treatment assignment at the next stage $k+1$ is independent of potential outcomes of the individual under any treatment options across all stages. Mathematically, we have $\forall a_k \in \mathcal{A}_k, A_k \perp\!\!\!\perp \{X_1, X_2^*(\bar{a}_1), \ldots, X_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K)\} \mid H_K$.

3. Positivity: we assume that for any tuples of history information of patient covariates and treatment assignments up to stage k that have a positive probability to be observed, the corresponding treatment regime will have a positive probability to be observed. Mathematically, we have $\forall k$ if $P[H_k = (\bar{x}_k, \bar{a}_{k-1})] > 0$, then with probability 1 $P[A_k = a_k \mid H_k] > 0$.

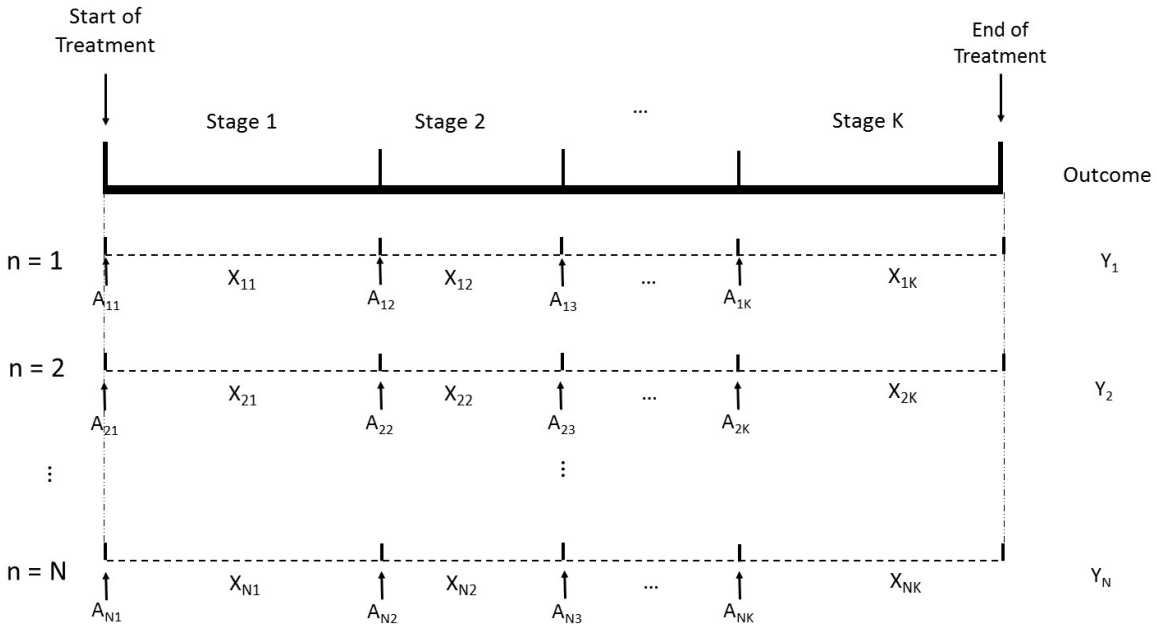

Figure 2.1: Standard data setting for dynamic treatment regimes

2.1.3 Observational Setting

Settings mentioned above are both experimental. However in many circumstances, observational data play a major role. For example, research on cancers and diabetes are not likely to involve human experimental data due to ethical issues, and thus studying on such diseases heavily rely on observational data. Electronic health record (EHR) provides insightful information for clinical diagnosis and clinical research but is also observational.

There are multiple benefits of using observational data (Kidwell 2015). First, obtaining observational data is usually the first step to understand disease characteristics and treatment effects. Second, using observational data is less likely to be concerned with ethical problems, such as research on cancers. Moreover, observational data are generally cheaper to collect than experimental data. In addition, especially for rare disease, it is more feasible to use observational data considering the small number of patients.

In order to make causal inference from observational data, we need to check whether assumptions are satisfied. The sequential ignorability assumption cannot be checked since

in observational studies, we cannot guarantee that covariate history up to any stage k is fully available. Besides, statistical methods for adjusting for confounding may not be applicable to time-varying treatment. Furthermore, deriving unbiased estimates for observational data is subject to model specification, even though this assumption may be weakened if using doubly robust methods (Kidwell 2015).

Considering these issues with observational settings, in this thesis we only consider the simpler settings where the assumptions for standard statistical methods are nicely specified.

2.2 Dynamic Treatment Regimes (DTRs) with Additional Longitudinal Data

2.2.1 Regularly Spaced Data

We consider a setting similar to DTRs with K stages except that additional data are collected across each stage. In addition to the treatment stages in a DTR, we allow an optional pre-treatment stage before stage one, which we will call stage 0. No treatment assignment is made in stage 0, but we may include preliminary patient data before the first treatment collected through electronic health record or mHealth etc. At each stage $k = 0, 1, \ldots, K$, patient data are collected at M_k randomly selected time points. For $k \geq 1$, we use $m = 0$ to denote the time when covariates measured right after treatment A_{nk} is assigned to patient n. To simplify the notations, we also allow $m = 0$ in the pre-treatment stage, even though no treatment assignments are made. We assume that covariates of all patients are measured at the same time, i.e. for each k, we obtain information on covariates of each patient n at $m = 0, \ldots, M_k$.

We use a $p-$dimensional vector X_{nk}^m to represent the available information of patient n at time m in stage k. For patient n in stage k, the available covariate information is $\mathbf{X}_{nk} = (X_{nk}^0, \ldots, X_{nk}^{M_k})^T$. Data of all patients in stage k will be denoted as $\mathbf{X}_k = (\mathbf{X}_{1k}, \mathbf{X}_{2k}, \ldots, \mathbf{X}_{Nk})^T$. Let Y_n be the final clinical outcome which is coded so that larger values correspond to better outcomes. Then similar to the standard DTRs, we can represent each patient's information as $(\mathbf{X}_{n0}, \mathbf{X}_{n1}, A_{n1}, \mathbf{X}_{n2}, A_{n2}, \ldots, \mathbf{X}_{nK}, A_{nK}, Y_n)^T$ where each

tuple is an independent and identically distributed trajectory of $(X_0, X_1, A_1, X_2, A_2, \ldots, X_K, A_K, Y)$. Let $H_k = (X_0, X_1, A_1, \ldots, A_{k-1}, X_k) \in \mathcal{H}_k$ be the history information up to stage k with $H_1 = (X_0, X_1)$, and $\mathcal{D}_k$ be the collection of available treatments at stage k. A dynamic treatment regime (DTR) in this setting, similar to before, is a sequence of decision rules $\mathbf{d} = (d_1, \ldots, d_K)$ where each d_k is a map from $\mathcal{H}_k$ to $\mathcal{D}_k$. Our goal is again to find the optimal DTR d^{opt} that maximizes the expected average outcome if the rule is implemented by the entire population in the future.

To formalize the process through potential outcomes, we can use the same notations as in standard DTRs, except that x_0 is added to all history variables. In addition, we make the same three assumptions for this modified setting, i.e. SUTVA, sequential ignorability, and positivity.

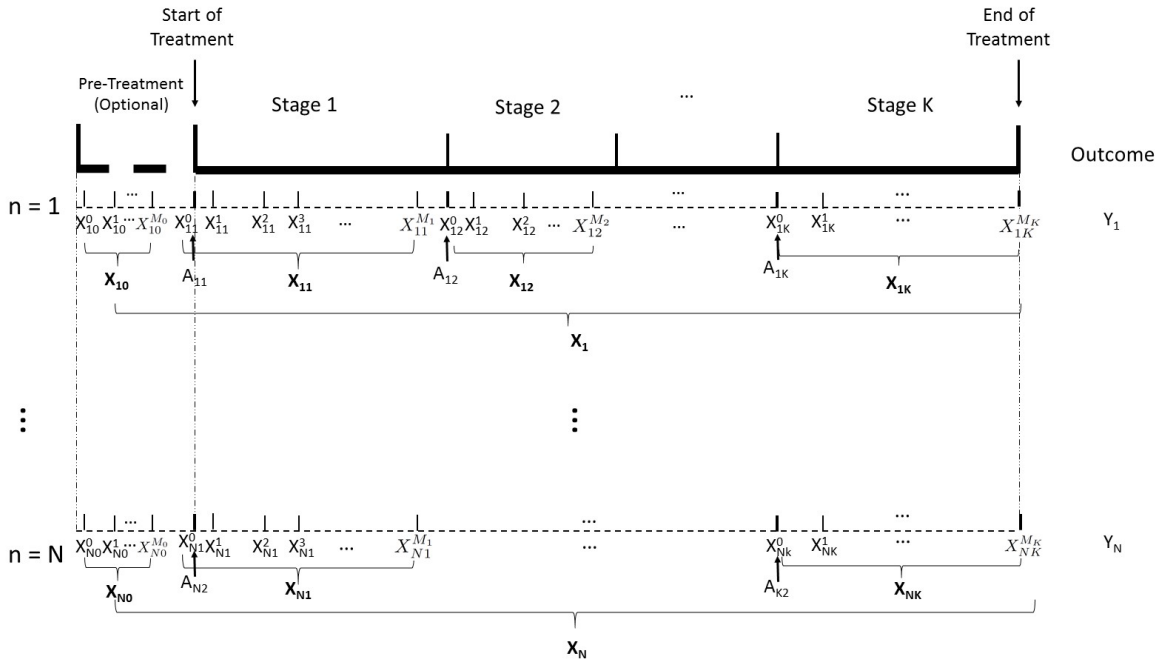

Figure 2.2: Data setting for dynamic treatment regimes with regularly spaced additional longitudinal data

2.2.2 Irregularly Spaced Data

In real clinical practice, patients covariates are rarely measured at regular gaps. Instead, nearly all patients' measuring time are different. Here we use irregularly spaced data to refer to the corresponding setting where at least one patient's covariates are observed at a different time than other patients' covariates, i.e. the data are asynchronous. In addition, given a stage, the number of observations for different patients may be different. Like the regularly spaced setting, we also allow stage 0, which is the optional pre-treatment stage before stage one. Suppose for patient n, we measure covariates at time $m_n = 0, 1, \ldots, M_{nk}$ at stage k. For $k \geq 1$, we use $m = 0$ to denote the time when covariates measured right after treatment A_{nk} is assigned to patient n. To simplify the notations, we also allow $m = 0$ in the pre-treatment stage, even though no treatment assignments are made.

As before, we use a $p-$dimensional vector $X_{nk}^{m_n}$ to represent the covariates of patient n at time m in stage k. For patient n in stage k, the available covariate information is $\mathbf{X}_{nk} = (X_{nk}^0, \ldots, X_{nk}^{M_{nk}})^T$. Data of all patients in stage k will be denoted as $\mathbf{X}_k = (\mathbf{X}_{1k}, \mathbf{X}_{2k}, \ldots, \mathbf{X}_{Nk})^T$. Let Y_n be the final clinical outcome which is coded so that larger values correspond to better outcomes. Then similar to the standard DTRs, we can represent each patient's information as $(\mathbf{X}_{n0}, \mathbf{X}_{n1}, A_{n1}, \mathbf{X}_{n2}, A_{n2}, \ldots, \mathbf{X}_{nK}, A_{nK}, Y_n)$ where each tuple is an independent and identically distributed trajectory of $(X_0, X_1, A_1, X_2, A_2, \ldots, X_K, A_K, Y)$. Although the proposed notations lead us to a similar situation to the previous one where data are regularly spaced, it is dangerous to apply this simple data representation to our analysis because the data are too sparse for most existing methods to be valid.

Considering the small amount of literature dealing with this complicated situation, which is closest to the reality, most of our attention in this thesis will be paid to the previous two multi-stage situations, where the data are not sparse or are regularly spaced for simplicity.

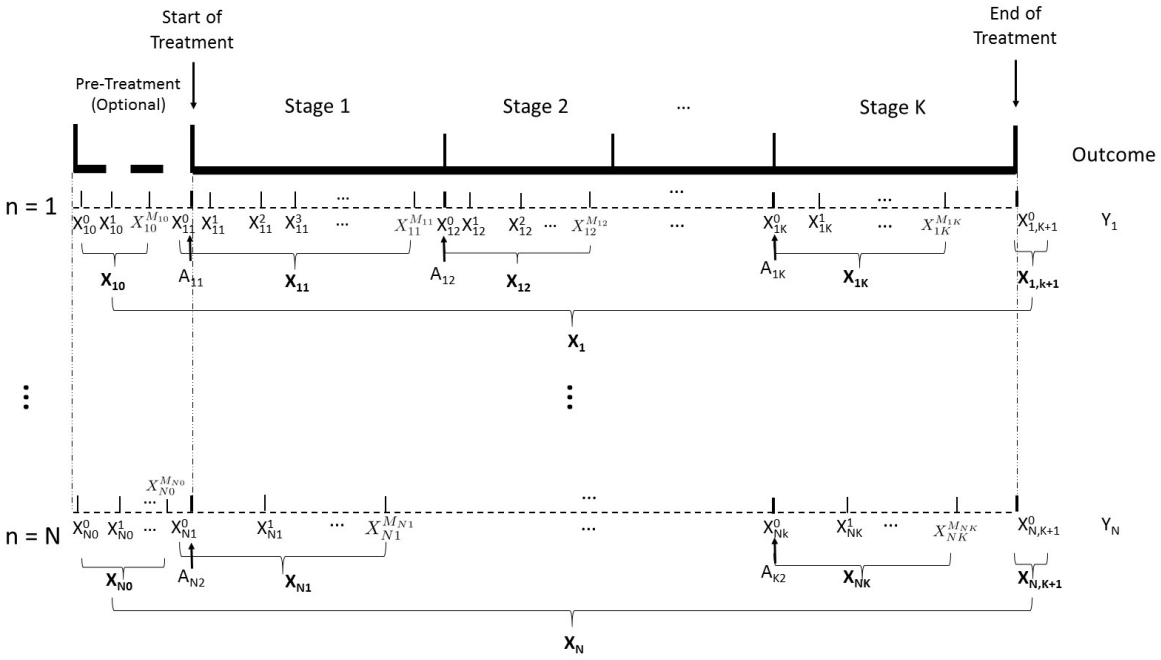

Figure 2.3: Data setting for dynamic treatment regimes with irregularly spaced additional longitudinal data

3 REINFORCEMENT LEARNING

We first review the statistical methods dealing with the standard DTRs. In treating life-threatening diseases such as breast cancer and lung cancer, many effective treatments involves multiple stages that are adaptive to patient performance. There are at least three challenges for statistical designs of adaptive treatment or trials. First, many existing designs are based on parametric models to account for efficacy, toxicity, and time to some events. For example, (Thall *et al.* 2000) provided a statistical framework for multi-stage treatment or clinical trials with modifications of the play-the-winner-and-drop-the-loser strategy, in which a successful treatment is repeated while an unsuccessful one is replaced by another treatment. Second, as a result of the parametric models, the heterogeneity in treatment across individuals is ignored and the heterogeneity needed for optimizing individualized treatment rule is not incorporated. Third, long-term benefits of the treatment or trials are not successfully evaluated due to delayed effects. Considering these challenges, the authors of (Zhao *et al.* 2009) presented a general reinforcement learning framework and related statistical methods for discovering new treatment regimes.

3.1 Reinforcement Learning and Q-Learning Backgrounds

The basic idea of reinforcement learning is to maximize the outcomes, called the rewards in this context, by telling the learning agent whether an action is "good" or "bad" when it tries among all available actions. In our DTR context, we use $\mathcal{X}$ and $\mathcal{A}$ to denote the space of patient covariates and the space of treatments respectively. In the reinforcement learning context, the random variables X and A are called "state" and "action" respectively. Define the time-dependent random variables states $\overline{X}_k = \{X_0, X_1, \ldots, X_k\}$ with realized values $\bar{x}_k = \{x_0, x_1, \ldots, x_k\}$, and actions $\overline{A}_k = \{A_0, A_1, \ldots, A_k\}$ with realized values

$\bar{a}_k = \{a_0, a_1, \ldots, a_k\}$. The state variables may or may not include past actions, i.e. X_k may include A_{k-1}. The distribution P from which the finite longitudinal trajectories are randomly sampled consists of the unknown distribution of each X_k conditional on previous $(\overline{X}_{k-1}, \overline{A}_{k-1})$ with conditional densities $\{f_0, \ldots, f_K\}$. The expectations of the conditional distributions with respect to the distribution P are denoted as E. For $k = 0, 1, \ldots, K$, we define the history information up to stage k for a patient as $H_k = (\overline{X}_k, \overline{A}_{k-1})$. Define the outcome of a patient's treatment after stage k as $Y_k = R(\overline{X}_k, \overline{A}_k, X_{k+1}) = R(H_{k+1})$ where R is a (possibly random) map from the space of states and actions to the space of real numbers. The realized value of the reward after stage k is $y_k = R(x_k, a_k, x_{k+1})$. Our goal is to find a_k to maximize the expected discounted return:

$$\tilde{y}_k = y_k + \gamma y_{k+1} + \gamma^2 y_{k+2} + \cdots + \gamma^K y_{k+K} = \sum_{i=0}^{K-k} \gamma^i y_{k+i}$$

where $\gamma \in [0, 1]$ is the discount rate. Intuitively, if γ is closer to 1, the future rewards are weighted more strongly.

A key element of the reinforcement learning framework is an exploration policy that maps past states and past actions to the probability that the next action a is taken given the past states and past actions, i.e. $p : h_k \mapsto p_t(a \mid h_k)$. We can write $d_k(h_k) = a_k$ if the policy is deterministic but not non-stationary where d_k is the decision rule in stage k. Denote the distribution of training data as P_d when the policy d is used to generate actions, and the corresponding expectations as E_d. The optimal sequence of treatment, or optimal policy here, maximizes the expectations with respect to the sum of the rewards over the time trajectories. We can represent our problem using a value function based on the state history h_k. The value function is the expected total future rewards of a patient conditional on h_k, i.e.

$$V_k(h_k) = E_d \left[\sum_{i=0}^{K-k} \gamma^i Y_{k+i} \mid H_k = h_k \right].$$

Then the optimal value function is

$$V_k^*(h_k) = \max_{d \in \mathcal{D}} V_k(h_k) = \max_{d \in \mathcal{D}} E_d \left[\sum_{i=0}^{K-k} \gamma^i Y_{k+i} \mid H_k = h_k \right].$$

In reinforcement learning, value functions are supposed to satisfy some recursive relationships. Therefore we can write the optimal policy d^{opt} as

$$d_k^{opt}(h_k) \in \operatorname*{argmax}_{a_k} E \left[Y_k + \gamma V_{k+1}^*(H_{k+1}) \mid H_k = h_k, A_k = a_k \right].$$

In reality, it is common that the optimal policy is not directly computable, and therefore the authors suggest using an alternative temporal-difference (TD) learning approach, specifically Q-learning which estimates a Q-function instead of the value function. Q-learning is an effective model-free algorithm that allows us to estimate the optimal strategies when we have insufficient knowledge about the distribution of the random variables. The optimal time-dependent Q-function is

$$Q_k^*(h_k, a_k) = E \left[Y_k + \gamma V_{k+1}^*(H_{k+1}) \mid H_k = h_k, A_k = a_k \right].$$

Since

$$V_k^*(h_k) = \max_{a_k} Q_k^*(h_k, a_k),$$

we have the optimal policy satisfying

$$d_k^{opt}(h_k) = \operatorname*{argmax}_{a_k} Q_k^*(h_k, a_k).$$

One-step Q-learning has the recursive form

$$Q_k(h_k, a_k) = E \left[Y_k + \gamma \max_{a_{k+1}} Q_{k+1}(H_{k+1}, a_{k+1}) \mid H_k = h_k, A_k = a_k \right]. \tag{3.1}$$

We let $\hat{Q}_k$ be the estimator of the optimal Q-functions for $k = 0, 1, \ldots, K$. According to (3.1), Q_k should be estimated backwards recursively from the last stage to the first stage. We can let $\hat{Q}_{K+1} = 0$ for convenience, and obtain $\hat{Q}_K$ first, then $\hat{Q}_{K-1}, \ldots, \hat{Q}_1, \hat{Q}_0$. Each Q_k can be viewed as a function of the states, actions, and a set of time-varying parameters θ, denoted as $\hat{Q}_k(h_k, a_k; \theta)$. Once we obtain the sequence of estimated Q-functions $\{\hat{Q}_0, \hat{Q}_1, \ldots, \hat{Q}_K\}$, we are able to estimate the optimal policies via

$$\hat{d}_k = \underset{a_k}{\operatorname{argmax}} \hat{Q}_k(h_k, a_k; \theta)$$

for $k = 0, 1, \ldots, K$.

3.2 Estimating the Q-Function

Fitting the Q functions has quite a few challenges. For example, the optimization problem in (3.1) is not smooth. The dimension of the state variables may be high. Action variables may also be of high dimension or even continuous. To deal with the difficulties, the authors presented two methods, support vector regression (SVR) and extremely randomized trees (ERT), for fitting Q-functions and learning the optimal policies.

3.2.1 Support Vector Regression

SVR is a flexible approach for regression problems, and the basic ideas of SVR are similar to those of SVM. To fit the settings of SVR into a more familiar framework, we denote the given training data $\{(z_n, y_n) \in \Omega \times \mathbb{R} : n = 1, \ldots, N\}$, where $\Omega = \{X, A : X \in \mathcal{X}, A \in \mathcal{A}\}$ and $\mathbb{R}$ is the real line representing the set of numerical rewards. We define the attributes $\mathbf{z}_{nk} \in \overline{X}_k \times \overline{A}_k$ for each $n = 1, \ldots, N$ and $k = 0, 1, \ldots, K$. Each total future numerical outcome y_{nk}.

To guarantee the data are separable when the dimension grows high, the data $\mathbf{z}_n$ are first mapped by a non-linear transformation Φ into the feature space. The Q function acts similarly to a hyperplane $f(\mathbf{z})$ that is fitted to the mapped data. We first suppose the function f is linear. Let $f(\mathbf{z}) = \mathbf{w}^T \Phi(\mathbf{z_n}) + b$ and the ε-insensitive loss function

$L(f(\mathbf{z_n}, y_n) = (|f(\mathbf{z_n}) - y_n| - \varepsilon)_+, \varepsilon > 0$. Other loss functions may also be appropriate. Then SVR solves the following optimization problem:

$$\min_{w,b,\xi,\xi'} \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} (\xi_n + \xi'_n),$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{w}^T \Phi(\mathbf{z_n}) + b - y_n \leq \varepsilon + \xi_n, \\ & y_n - \mathbf{w}^T \Phi(\mathbf{z_n}) + b \leq \varepsilon + \xi'_n, \\ & \xi_n, \xi'_n \geq 0, n = 1, \ldots, N, \end{aligned} \tag{3.2}$$

where ξ_n and ξ'_n are slack variables, and C is the cost of error, also called the tuning parameter. The goal of the above setup is to discover a function that has at most ε deviation from the actual values y_n for all training data.

The authors also provide a framework for non-linear kernels. Kernels are a class of non-negative functions that measure the similarity between features of the individuals, requiring no knowledge of the non-linear transformation. The kernel function $K : \Omega \times \Omega \to \mathbb{R}$ is continuous, symmetric, and positive definite. We can associate with it a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ which is the completion of the linear span of all functions $\{K(\cdot, \mathbf{z}) : \mathbf{z} \in \Omega\}$, with norm induced by the inner product. So we can define $K(\mathbf{z_i}, \mathbf{z_j}) = \Phi(\mathbf{z_i})^T \Phi(\mathbf{z_j})$. Equation (3.2) can be rewritten as

$$\min_{\lambda,\lambda'} \frac{1}{2} (\lambda - \lambda')^T K(z_i, z_j)(\lambda - \lambda') + \varepsilon \sum_{i=1}^{N} (\lambda - \lambda') + \sum_{i=1}^{N} y_i (\lambda - \lambda')$$

$$\text{subject to} \sum_{i=1}^{N} (\lambda - \lambda') = 0, 0 \leq \lambda_i, \lambda'_i \leq C, i = 1, \ldots, N.$$

Solving for the optimal λ and λ', we get the approximating function

$$f(z) = \sum_{i=1}^{N} (\lambda - \lambda') K(x_i, z) + b.$$

3.2.2 Extremely Randomized Trees

The other method to estimate the Q-function mentioned by the authors is the ERT, which was originally proposed by (Ernst *et al.* 2005). This nonparametric method uses random forests and builds each tree by randomizing both attribute and cut-point choice when splitting a tree node. The parameters include the number of trees, the maximum number of cut-direction tests at each node, and the minimum number of elements to split a leaf. See (Ernst *et al.* 2005) for more details about the algorithm.

3.3 Discussion

To demonstrate the use and effectiveness of the proposed methods, the authors apply the methods to a simulated sequential multiple assignment randomized trial (SMART). The result shows that the Q-learning approach using either SVR or ERT performs better in discovering the optimal policy with roughly equal computational costs.

On discrete state spaces, the proposed method can be applied to any type of treatment, including for example, continuous dose ranges and binary options. It does not depend on specific accurate mathematical models. It takes into consideration drug efficacy and toxicity simultaneously and improves the long-term outcomes. The method is also applicable to high-dimensional attributes with relatively low computational burden. On the other hand, some potential future improvements include for example, robustness to the model of the reward function, incorporation of patient and physician preference, and addressing reversible toxicity of the drug.

4 OUTCOME WEIGHTED LEARNING

A common approach to estimate individualized treatment rules with binary treatment options in a one-stage setting is regression. However, most regression-based methods are parametric or semi-parametric to estimate and optimize the conditional means. (Zhao *et al.* 2012) proposed a new method to avoid modeling the conditional means, but to estimate directly the decision rule that maximizes clinical response.

Using the same notations as in previous sections, the proposed method applies to binary treatment assignments $A \in \mathcal{A} = \{-1, 1\}$, and each patient's prognostic variables are $X = (X_1, \ldots, X_p)^T \in \mathcal{X}$. We assume the reward R is bounded and is coded so that larger values correspond to better clinical outcomes. The optimal ITR is the rule that maximizes the expected reward if implemented by the entire population. We let the distribution of (X, A, R) be P and the expectation with respect to P be E. Given any treatment rule d, we denote the distribution of (X, A, Y) as P^d and the expectation with respect to P^d as E^d.

Under the assumption of positivity, i.e. $P(A = a) > 0$ for $A = -1$ and 1, we have P^d being absolutely continuous with respect to P and

$$\frac{dP^d}{dP} = \frac{I\{a = d(x)\}}{P(A = a)}$$

Thus the expected reward under the rule d, also called the value function associated with d, is

$$\mathcal{V}(d) \doteq E^d(Y) = \int Y dP^d = \int R \frac{dP^d}{dP} dP = E\left[\frac{I\{A = d(X)\}}{A\pi + (1 - A)/2} Y\right]$$

where $\pi = P(A = 1)$. As a result, the optimal ITR is

$$d^{opt} \in \underset{d}{\operatorname{argmax}} E\left[\frac{I\{A = d(X)\}}{A\pi + (1-A)/2} Y\right].$$

Since we can see in the above formula that the right hand side is location invariant in Y, we may assume Y is nonnegative with out loss of generality.

To estimate the optimal ITR, we can equivalently find

$$d^{opt} \in \underset{d}{\operatorname{argmin}} E\left[\frac{I\{A \neq d(X)\}}{A\pi + (1-A)/2} Y\right] \tag{4.1}$$

since Y is assumed to be bounded. The right hand side of (4.1) can be viewed as minimizing a weighted classification error, where each misclassified A using X is weighted by $\frac{Y}{A\pi+(1-A)/2}$. We would like to find a decision function f such that $d(x) = \text{sign}\{f(x)\}$ with $I\{a = d(x)\} = I\{af(x) > 0\}$. We can therefore approximate (4.1) by the empirical value

$$\mathbb{P}_N\left[\frac{I\{A \neq \text{sign}\{f(X)\}\}}{A\pi + (1-A)/2} Y\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{Y_i}{A_i\pi + (1-A_i)/2} I\{A \neq \text{sign}\{f(X_i)\}\} \tag{4.2}$$

where $\mathbb{P}_N$ denotes the empirical measure of the observed data. However, equation (4.2) involves minimizing a discontinuous and nonconvex 0-1 loss. One common solution is to use a surrogate loss function, such as the hinge loss. Then to minimize equation (4.2), we can instead minimize

$$\frac{1}{N}\sum_{i=1}^{N} \frac{Y_i}{A_i\pi + (1-A_i)/2} (1 - A_i f(X_i))^+ + \lambda_N \|f\|^2 \tag{4.3}$$

where $x^+ = max(x, 0)$ and $\|f\|$ is some norm of f.

Linear Decision Rule for Optimal ITR Suppose f is a linear decision function with $f(x) = \boldsymbol{\beta}^T x + \beta_0$. The corresponding decision rule to assign a patient with prognostic

value X to treatment 1 is $\boldsymbol{\beta}^T x + \beta_0 > 0$ and -1 otherwise. We let the norm in equation (4.3) be the Euclidean norm. Following the usual SVM, we can rewrite equation (4.3) as

$$\begin{aligned} & \max_{\boldsymbol{\beta},\beta_0,\|\boldsymbol{\beta}\|=1} C \\ \text{subject to} \quad & A_i(\boldsymbol{\beta}^T x + \beta_0) \geq C(1-\xi_i), \\ & \xi_i \geq 0, \sum \frac{R_i}{\pi_i}\xi_i < s, \end{aligned} \tag{4.4}$$

where $C > 0$ is the classifier margin, $\pi_i = \pi I\{A_i = 1\} + (1-\pi)I\{A_i = -1\} = P(A = 1|X_i)$ and s is a constant depending on λ_N. Note that equation (4.4) is equivalent to

$$\begin{aligned} & \min \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \kappa \sum_{i=1}^{N} \frac{R_i}{\pi_i}\xi_i \\ \text{subject to} \quad & A_i(\boldsymbol{\beta}^T x + \beta_0) \geq (1-\xi_i), \\ & \xi_i \geq 0, \end{aligned}$$

where κ is a tuning parameter. After introducing Lagrange multipliers and algebraic manipulations, we obtain a dual problem

$$\begin{aligned} & \max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j A_i A_j X_i^T X_j \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \kappa R_i/\pi_i, i = 1, \ldots, N, \\ & \sum_{i=1}^{N} \alpha_i A_i = 0. \end{aligned} \tag{4.5}$$

This dual problem involves a quadratic objective function. Finally we obtain

$$\hat{\boldsymbol{\beta}} = \sum_{\hat{\alpha}_i > 0} \hat{\alpha}_i A_i X_i,$$

and estimate $\hat{\beta}_0$ using the marginal points $(0 < \hat{\alpha}_i, \hat{\xi}_i = 0)$.

Nonlinear Decision Rule for Optimal ITR In most cases the decision rule is likely to be nonlinear due to the complicated structure of the space of prognostic variables. We use the kernel function K, as introduced in section 3.2.1, to find the decision function f. Since $f(x)$ comes from the associated RKHS $\mathcal{H}_K$, it can be written as a linear combination of $K(\cdot, x)$, i.e. $f(x) = \sum_{i=1}^{m} \alpha_i K(\cdot, x_i)$. We can show that the optimal decision function is given by

$$\sum_{i=1}^{N} \hat{\alpha}_i A_i K(X, X_i) + \hat{\beta}_0,$$

where $(\hat{\alpha}_1, \ldots, \hat{\alpha}_N)$ solves the same dual problem as in equation (4.5)

The authors also establish several properties of the optimal ITR estimated by OWL. First, the risk associated with the optimal decision rule under 0-1 loss is the Bayes risk. Second, Fisher consistency is established to justify the validity of using the surrogate loss function, hinge loss, in OWL. Third, the excess risk of f under 0-1 loss is no larger than the risk of f under hinge loss. Fourth, the value of the estimated optimal decision function $\hat{f}_N$ is a consistent estimator of the true optimal value function.

OWL is the first to introduce machine learning methods that directly estimate the optimal ITR without appealing to any accurate mathematical models of the conditional means. It also avoids overfitting compared to other two-stage methods. Furthermore, the convergence rate of OWL is nearly the optimal for nonparametric SVM on completely separated data.

5 BACKWARD AND SIMULTANEOUS OUTCOME WEIGHTED LEARNING

In previous sections, we introduced Q-learning for estimating the optimal dynamic treatment regime. It estimates the Q-functions using the data first, and then maximizes or minimizes the function to infer the optimal DTRs. However, this two-step regression-based method encounters some issues when facing high-dimensional data. To resolve such issues, (Zhao *et al.* 2015) proposed two new dynamic statistical learning approaches to estimating the optimal DTR. One method is called backward outcome weighted learning (BOWL), which treats optimal DTR estimation as a sequence of weighted classification problems. It uses outcome weighted learning to identify a sequence of optimal decision rules in a backward recursive fashion. The other method is called simultaneous outcome weighted learning (SOWL), which treats optimal STR estimation as a single classification problem. It uses outcome weighted learning to identify the optimal decision rules at all stages simultaneously.

5.1 Backward Outcome Weighted Learning (BOWL)

We suppose the three assumptions for DTRs with experimental data hold: causal consistency, sequential ignorability, and positivity. The treatment option is binary $A_k \in \mathcal{A} = \{-1, 1\}$. Covariate history up to stage k is denoted with an overbar $\overline{X}_k = \{X_0, X_1, \ldots, X_k\}$, with realized values $\bar{x}_k = \{x_0, x_1, \ldots, x_k\}$. The actions taken up to stage k is denoted as $\overline{A}_k = \{A_0, A_1, \ldots, A_k\}$ with realized values $\bar{a}_k = \{a_0, a_1, \ldots, a_k\}$. Similar to outcome weighted learning in the previous chapter, we would like to maximize the value associated

with a decision rule d, which is

$$\mathcal{V}(d) = \int Y \frac{dP^d}{dP} dP = E\left[\frac{Y \prod_{j=k+1}^{K} I\{A_j = d_j(H_j)\}}{\prod_{j=k}^{K} \pi_j(H_j, A_j)}\right]. \tag{5.1}$$

The idea behind BOWL is through backward estimation based on future optimal decision rules that are available. Suppose that we have obtained all optimal treatment rules after stage k, denoted as $\underline{d}_{k+1}^{opt} = (d_{k+1}^{opt}, \ldots, d_K^{opt})$. Then the optimal decision rule at stage k should maximize

$$E\left[\frac{Y \prod_{j=k+1}^{K} I\{A_j = d_j^{opt}(H_j)\}}{\prod_{j=k}^{K} \pi_j(H_j, A_j)} I\{A_k = d_k(H_k)\} \,\middle|\, H_k = h_k\right].$$

This places the constraint that we only consider patients who follow exactly the optimal treatment regime in all stages after the k-th stage. Equivalently, the optimal treatment regime d^{opt} is a map from $\mathcal{H}_k$ to $\{-1, 1\}$ that minimizes the empirical analogue of the above expression:

$$E\left[\frac{Y \prod_{j=k+1}^{K} I\{A_j = d_j^{opt}(H_j)\}}{\prod_{j=k}^{K} \pi_j(H_j, A_j)} I\{A_k \neq d_k(H_k)\}\right]. \tag{5.2}$$

This can be viewed as an optimization problem with 0-1 loss or a weighted misclassification problem, where the weights are defined by

$$\frac{Y \prod_{j=k+1}^{K} I\{A_j = d_j^{opt}(H_j)\}}{\prod_{j=k}^{K} \pi_j(H_j, A_j)}.$$

To develop an estimation procedure, the authors replace the 0-1 loss function with a convex surrogate loss function $\phi(t)$. Let $f_k : \mathcal{H}_j \to \mathbb{R}$ denote the decision function in stage k, so that $d_k(h_k) = \text{sign}(f_k(h_k))$. Then we can minimize the following expression with respect

to f_k:

$$\mathbb{P}_N \left[\frac{Y \prod_{j=k+1}^{K} I\{A_j = d_j^{opt}(H_j)\}}{\prod_{j=k}^{K} \pi_j (H_j, A_j)} \phi \left(A_k f_k(h_k) \right) \right] + \lambda_{k,N} \left\| f_k \right\|^2, \tag{5.3}$$

where $\lambda_{k,N}$ is a tuning parameter controlling the amount of penalization. Since we do not know the future optimal DTR, we need to estimate the decision function from the last stage and proceed backwards. The BOWL algorithm is presented as follows:

Algorithm 1 BOWL

Input: Patient history up to stage k : $H_k = (\overline{X}_k, \overline{A}_{k-1})$

Output: Decision rule d_k at stage k

for $k \leftarrow K, K-1, \ldots, 1$ **do**

  if $k = K$ **then** $\hat{f}_k \in \operatorname{argmin}_{f_k} \left\{ \mathbb{P}_N \left[\frac{Y}{\pi_k(H_k, A_k)} \phi \left(A_k f_k(h_k) \right) \right] + \lambda_{k,N} \left\| f_k \right\|^2 \right\}$

    $\hat{d}_k(h_k) \leftarrow \text{sign}(\hat{f}_k(h_k))$

  else $\hat{f}_k \in \operatorname{argmin}_{f_k} \left\{ \mathbb{P}_N \left[\frac{Y \prod_{j=k+1}^{K} I\{A_j = \hat{d}_j(H_j)\}}{\prod_{j=k}^{K} \pi_j(H_j, A_j)} \phi \left(A_k f_k(h_k) \right) \right] + \lambda_{k,N} \left\| f_k \right\|^2 \right\}$

    $\hat{d}_k(h_k) \leftarrow \text{sign}(\hat{f}_k(h_k))$

  end if

end for

In the above algorithm, the minimization problem is similar to that in outcome weighted learning in the previous section.

5.2 Simultaneous Outcome Weighted Learning

Unlike BOWL which estimates the optimal decision rules sequentially, SOWL completes the task at all stages simultaneously. However, maximizing (5.1) involves a discontinuous, non-convex 0-1 loss function, which may cause computational complexity. For $k = 1, \ldots, K$, let $Z_k = A_k f_k(H_k)$. In SOWL, noticing that $\prod_{j=k+1}^{K} I\{A_j = \hat{d}_j(H_j)\}$ is equivalent to $\prod_{j=k+1}^{K} I\{Z_k > 0\}$, the authors replace the 0-1 loss function with hinge loss $\psi(Z_1, \ldots, Z_K) = \min(Z_1 - 1, \ldots, Z_K - 1, 0) + 1$, which is smooth and concave. Then

the objective function to maximize is

$$\mathbb{P}_N \left[\frac{Y\psi(Z_1, \ldots, Z_K)}{\prod_{j=k}^{K} \pi_j \left(H_j, A_j\right)} \right] - \lambda_N \sum_{k=1}^{K} \|f_k\|^2 , \tag{5.4}$$

where λ_N is a tuning parameter controlling the amount of penalization. The detailed computational algorithm is presented in (Zhao *et al.* 2015) and we will omit the details here.

5.3 Discussion

Both BOWL and SOWL aim at maximizing directly the expected long-term outcome. Compared to regression-based Q-learning, they are more robust because they do not relying on models of the Q-function for the optimal DTRs. When the number of stages is large, SOWL may face som numerical instability because is involves defining a multi-dimensional surrogate loss. In this case, BOWL may have more benefits. On the other hand, SOWL allows one to examine the estimation at the same time rather than sequentially using BOWL.

6 PROPOSED METHODOLOGY FOR ADDITIONAL LONGITUDINAL DATA

The main purpose of this section is to provide a genuine framework for estimating the optimal individualized treatment regime when sparse asynchronous longitudinal data are present. Here we only consider binary treatment options $A \in \mathcal{A} = \{-1, 1\}$ in a randomized study. We assume that patient covariates are time-varying and can be viewed as a function of time, i.e. $X = X(t)$, and therefore patient history can also be viewed as a time-varying variable $H_k(t) = (X(t)_0, X(t)_1, A_1, \ldots, A_{k-1}, X(t)_k) \in \mathcal{H}_k$. To define the optimal ITR, we make the same three assumptions for dynamic treatment regimes: (1) causal consistency; (2) sequential ignorability; and (3) positivity. Define

$$Q(h, a) = E[Y \mid H(t) = h(t), A = a].$$

To model on the conditional mean, we will fit a linear model with the intercept of the form

$$E\left[Y \mid H(t) = h(t), A = a\right] = \alpha + \beta^T h(t) + a\{\gamma + \delta^T h(t)\}, \tag{6.1}$$

where α and β are unknown regression parameters of the main effects, and γ and δ are unknown regression parameters of the interaction effects.

The basic idea behind our methodology is to apply a counting process approach to generate new features using a set of basis functions in each stage, and the optimal treatment rule is estimated with the new features by a Q-learning-like procedure. We use a counting process for the observation times of the covariates to represent the asynchronous data of

our setting. In stage k for each patient $n = 1, \ldots, N$, define

$$F_{nk}(t) \doteq \sum_{j=1}^{M_{nk}} I\{T_{nk}^{j} \leq t\} \tag{6.2}$$

where $T_{nk}^{j}, j = 1, \ldots, M_{nk}$ are the observation times for the covariates in stage k and $M_{nk} < \infty$ with probability 1. Thus the actual observations on the covariates are $X(T_{nk}^{1}), \ldots, X(T_{nk}^{M_n k})$.

For each stage $k = 1, \ldots, K$, let $\Psi_k = \{\Phi_{kl}(t; \theta_{kl}) : \theta_{kl} \in \Theta, l = 1, \ldots, L_k\}$ be a collection of normalized basis functions chosen to model patient covariates $X(t)$. Each $\Phi_{kl}(\cdot; \theta_{kl})$ is indexed by an unknown parameter θ_{kl}, either a vector or a scalar, in the parameter space Θ. We require each $\Phi_{kl}(\cdot; \theta_{kl})$ be of the form $\frac{\phi(\cdot; \theta_{kl})}{c(\theta_{kl})}$ where $c(\theta_{kl})$ is a constant that averages the effect of the chosen basis function on the covariates over the entire time period with respect to the counting process. If the basis function is a kernel function, then $c_{kl}(\theta_{kl}) = \int \Phi_{kl}(t; \theta_{kl}) dF_{nk}(t)$. Take the radial basis function kernel $\Phi_{RBF}(t; \theta) = c(\theta) exp(-\theta \left\| t - t' \right\|)^2$ as an example, which downweights the observations made distant in time to a given time point t'. Assuming time $t \geq 0$, $c(\theta) = \int_0^\infty exp(-\theta \left\| t - t' \right\|)^2 dt = \sqrt{\frac{\pi}{\theta}} \Phi(\sqrt{2\theta} t')$ where Φ here is the standard normal cumulative distribution function. On the other hand, if the basis function is a natural basis, then $c_{kl}(\theta_{kl}) = \int dF_{nk}(t)$. These bases include $\cos(\theta t), \sin(\theta t)$, and $a + \theta t$, etc. The basis functions are to be chosen by the investigator depending on the research questions of interest as well as data structure.

We now begin to construct new features for each patient's covariates using L_k basis functions. We view the new feature as a function of the parameter θ_l by defining

$$\tilde{X}_{nk}^{l}(\theta_{kl}) \doteq \int \Phi_{kl}(t; \theta_{kl}) X_{nk}(t) dF_{nk}(t) \tag{6.3}$$

with appropriate choice of bandwidth. The intuition behind the above definition is that given a time t, each new feature contains more information about the covariate observations

made before and close in time to t, than those made after or distant in time to t. This allows for using all covariate information we observe but focusing on time points of more interest. Let $\theta_k = (\theta_{k1}, \ldots, \theta_{kL_k})$. We use

$$U_{nk}(\theta_k) = (\tilde{X}^1_{nk}(\theta_1), \ldots, \tilde{X}^{L_k}_{nk}(\theta_{L_k}))^T$$

to denote the vector of new features of patient n in stage k, generated using the collection of kernels in Ψ. We also denote the collection of new features of all patients in stage k by $U_k(\theta) = (U_{1k}(\theta), \ldots, U_{Nk}(\theta))^T$. As before, we use an overbar to denote events that happened in the past, so $\overline{U}_k(\theta) = (U_1(\theta_1), \ldots, U_k(\theta_{L_k}))$. The history data up to stage k with the new features is then denoted as $\tilde{H}_k(\theta) = (\overline{U}_k(\theta), \overline{A}_{k-1})$.

As in reinforcement learning which was described in Chapter 3, One-step Q-learning has the recursive form

$$Q_k(\tilde{h}_k(\theta), a_k) = E\left[Y_k + \max_{a_{k+1}} Q_{k+1}(\tilde{H}_{k+1}(\theta), a_{k+1}) \mid \tilde{H}_k(\theta) = \tilde{h}_k(\theta), A_k = a_k\right], \quad (6.4)$$

where $Y_k = Y$ if $k = K$ and $Y_k = 0$ otherwise in our case.

We let $\hat{Q}_k$ be the estimator of the optimal Q-functions for $k = 0, 1, \ldots, K$. According to (6.4), Q_k should be estimated backwards recursively from the last stage to the first stage. We can let $\hat{Q}_{K+1} = 0$ for convenience, and obtain $\hat{Q}_K$ first, then $\hat{Q}_{K-1}, \ldots, \hat{Q}_1, \hat{Q}_0$. Note that the estimated Q-values are functions of the unknown parameter $\mu = (\alpha, \beta, \gamma, \delta, \theta)^T$, and we would also like to obtain estimates $\hat{\mu}_k = (\hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k, \hat{\delta}_k, \hat{\theta}_k)$ of the parameter. We may estimate the parameter using least-square simultaneously when estimating the optimal policy, so then

$$\hat{\mu}_k \in \underset{\mu}{\operatorname{argmin}}\, \mathbb{P}_N \left[Y_k + \max_{a_{k+1}} \hat{Q}_{k+1}(\tilde{H}_{k+1}(\theta), a_{k+1}; \hat{\mu}_{k+1}) - Q_k(\tilde{h}_k(\theta), a_k; \mu)\right]^2 \quad (6.5)$$

Notice that

$$\begin{aligned}
&\max_{a_{k+1}} \hat{Q}_{k+1}(\tilde{H}_{k+1}(\theta), a_{k+1}; \hat{\mu}_{k+1}) \\
=&\max_{a_{k+1}} \left[\alpha + \beta^T \tilde{h}_{k+1}(\theta) + a\{\gamma + \delta^T \tilde{h}_{k+1}(\theta)\}\right] \\
=&\alpha + \beta^T \tilde{h}_{k+1}(\theta) + \left|\gamma + \delta^T \tilde{h}_{k+1}(\theta)\right|.
\end{aligned}$$

Plugging in the linear model, we can rewrite the estimating procedures in (6.5) as:
When $k = K+1, \min_\mu \mathbb{P}_N \left(Y - \alpha - \beta^T \tilde{h}_k(\theta) - a_k\{\gamma + \delta^T \tilde{h}_k(\theta)\}\right)^2$;
When $k \leq K$,

$$\begin{aligned}
\min_\mu \mathbb{P}_N \Big(&\hat{\alpha}_k + \hat{\beta}_k^T \tilde{h}_k(\theta) + \left|\hat{\gamma}_k + \hat{\delta}_k^T \tilde{h}_k(\theta)\right| \\
&-\alpha_{k-1} - \beta^T \tilde{h}_{k-1}(\theta) - a_{k-1}\{\gamma_{k-1} + \delta_{k-1}^T \tilde{h}_{k-1}(\theta)\}\Big)^2.
\end{aligned} \tag{6.6}$$

After obtaining parameter estimates $\hat{\mu}_k$ for all $k = 1, \ldots, K$, the optimal decision rule for each stage can be estimated by

$$\hat{d}_k\left(\tilde{H}_k = \tilde{h}_k\right) = \text{sign}\left(\hat{\gamma}_k + \hat{\delta}_k^T \tilde{h}_k(\hat{\theta}_k)\right).$$

In order to determine our basis functions, the parameter θ is required to be identifiable, which means the one-to-one correspondence between the parameter and the basis function. Following our notation, we can express identifiability as $\Phi(t;\theta) = \Phi(t;\theta_0)\ \forall t$ if and only if $\theta = \theta_0$. In some cases, the parameter θ may not be identifiable if we do not restrict the range of the parameter space. Consider the example of a linear basis $\Phi(t; a, b) = a + bt$. To construct a new feature by (6.3), we calculate $\tilde{X} = \int (a + bt) X(t) dF(t)$. However, when estimating the parameters using (6.6), there may be multiple solutions to (a, b). One way to resolve this issue is to restrict the solution (a, b) to the unit circle using parametrization $a = \sin(\theta), b = \cos(\theta)$, where $\theta \in [0, 2\pi)$. The actual way of ensuring identifiability depends upon the form of the chosen basis functions.

7 DISCUSSION

In this thesis, we introduced individualized treatment regimes in single-stage and multi-stage settings, and a new setting where each patient's data are not collected at the same time. We described three types of recently developed machine learning method to estimate the optimal treatment rules in single-stage settings and dynamic treatment regimes. First, we introduced a regression-based nonparametric reinforcement learning method, Q-learning, which estimates the optimal decision rule by directly maximizing the Q-function sequentially backwards. Second, we reviewed outcome weighted learning for estimating individualized treatment rule in single-stage settings with binary treatment options. Third, we presented two nonparametric extensions of outcome weighted learning to dynamic treatment regimes. Backward outcome weighted learning and simultaneous outcome weighted learning estimate the optimal DTR directly by maximizing the long-term expected outcome over all possible DTRs.

We proposed a novel machine learning method to estimate the optimal DTR using sparse asynchronous data from a sequential randomized trial when the treatment option is binary. The method effectively deals with sparsity in asynchronous data by employing a counting process on patient covariates, and constructing new features with basis functions. It can handle the potentially complex structure in autocorrelation of covariates. We postulated a linear model on the expected long-term outcome given covariate history and treatment history, and applied Q-learning for estimation. Parameter estimation was completed through sequential backward least-square regression and the optimal decision was determined by the sign of the decision function.

However, our method may be subject to model misspecification, which may affect consistency of the estimates and stability of the method. Such misspecification may come from choices of basis functions and modeling of the conditional mean. Some of the concerns with potential model misspecification can potentially be addressed through extending the proposed longitudinal data approach to outcome weighted learning, including BOWL and SOWL, which is a good topic for future research. In addition, we assume the data we have are experimental. Applying this method to observational data may require modifying some assumptions on causal relationships of the covariates. Moreover, we also expect some future work to test the proposed method on simulated data and real data. Extending the treatment option from binary to multiple discrete options and continuous options is also a promising direction for future research.

8 REFERENCES

Anderbro, T., Amsberg, S., Adamson, U., Bolinder, J., Lins, P. E., Wredling, R., Moberg, E., Lisspers, J. and Johansson, U. B. (2010) Fear of hypoglycaemia in adults with Type 1 diabetes. *Diabetic Medicine*, **27**, 1151–1158.

Cao, H., Donglin, Z. and Fine, J. P. (2015) Regression analysis of sparse asynchronous longitudinal data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **77**, 755776.

Ernst, D., Geurts, P. and Wehenkel, L. (2005) Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, **6**, 503–556. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.7705{\&}amp;rep=rep1{\&}amp;type=pdf.

Kidwell, K. M. (2015) DTRs and SMARTs : Definitions , designs , and applications. In *Adaptive Treatment Strategies in Practice Planning Trials and Analyzing Data for Personalized Medicine* (eds. M. Kosorok and E. Moodie), chap. 2, 11–12.

Liu, L. L., Lawrence, J. M., Davis, C., Liese, A. D., Pettitt, D. J., Pihoker, C., Dabelea, D., Hamman, R., Waitzfelder, B. and Kahn, H. S. (2010) Prevalence of overweight and obesity in youth with diabetes in USA: The SEARCH for Diabetes in Youth Study. *Pediatric Diabetes*, **11**, 4–11.

McClure, J., Anderson, M., Bradley, K., An, L. and Sheryl, C. (2016) Evaluating an Adaptive and Interactive mHealth Smoking Cessation and Medication Adherence Program: A Randomized Pilot Feasibility Study. *JMIR mHealth and uHealth*, **4**, e94.

Murphy, S. A. (2005) An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, **24**, 1455–1481.

Murphy, S. A., Oslin, D. W., Rush, A. J. and Zhu, J. (2006) Methodological Challenges in Constructing Effective Treatment Sequences for Chronic Psychiatric Disorders. *Neuropsychopharmacology*, 257 262.

Pinhas-Hamiel, O. and Levy-Shraga, Y. (2013) Eating disorders in adolescents with type 2 and type 1 diabetes. *Current Diabetes Reports*, **13**, 289–297.

Schober, E., Wagner, G., Berger, G., Gerber, D., Mengl, M., Sonnenstatter, S., Barrientos, I., Rami, B., Karwautz, A. and Fritsch, M. (2011) Prevalence of intentional under- and overdosing of insulin in children and adolescents with type 1 diabetes. *Pediatric Diabetes*, **12**, 627–631.

Thall, P. F., Millikan, R. E. and Sung, H. G. (2000) Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, **19**, 1011–1028.

Wadden, T. a., Brownell, K. D. and Foster, G. D. (2002) Obesity: responding to the global epidemic. *Journal of consulting and clinical psychology*, **70**, 510–525.

Zhao, Y., Kosorok, M. R. and Zeng, D. (2009) Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, **28**, 3294–3315. URL http://dx.doi.org/10.1002/sim.3720.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012) Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **107**, 499–1106. URL http://www.tandfonline.com/loi/uasa20http://dx.doi.org/10.1080/01621459.2012.695674http://www.tandfonline.com/.

Zhao, Y.-Q. (2015) Outcome weighted learning methods for optimal dynamic treatment regimes. In *Adaptive Treatment Strategies in Practice Planning Trials and Analyzing Data for Personalized Medicine* (eds. M. R. Kosorok and E. E. M. Moodie), chap. 8, 127–129.

Zhao, Y.-Q., Zeng, D., Laber, E. B. and Kosorok, M. R. (2015) New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association*, **110**, 583–592.