Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings

Ashkan Ertefaie

University of Michigan

ertefaie@umich.edu

June 4, 2014

Abstract

The application of existing methods for constructing optimal dynamic treatment regimes is limited to cases where investigators are interested in optimizing a utility function over a fixed period of time (finite horizon). In this manuscript, we develop an inferential procedure based on temporal difference residuals for optimal dynamic treatment regimes in infinite-horizon settings where there is no a priori fixed end of follow-up point. The proposed method can be used to determine the optimal regime in chronic diseases where patients are monitored and treated throughout their life. We derive large sample results necessary for conducting inference. We also simulate a cohort of patients with diabetes, which mimics the third wave of the National Health and Nutrition Examination Survey, and we examine the performance of the proposed method in controlling the level of hemoglobin A1c.

Keywords: Action-value function; Causal inference; Backward induction; Temporal Difference residual.
1. INTRODUCTION

A dynamic treatment regime (DTR) is a treatment process that considers patients’ individual characteristics and their ongoing performance to decide which treatment option to assign. DTRs can, potentially, reduce possible side effects and treatment costs, which makes the process attractive for policy makers. The optimal DTR is the one that, if followed, yields the most favorable outcome on average. Depending on the context, the DTR is also called an adaptive intervention (Collins et al., 2004) or adaptive strategies (Lavori and Dawson, 2000).

The goal of this manuscript is to devise a new methodology that can be used to construct the optimal dynamic treatment regime in infinite-horizon settings (i.e., when the number of decision points is not necessarily fixed for all the individuals). The estimation procedure, however, is based on observational data collected over a fixed period of time which includes many decision points. One potential application of our method is to estimate the optimal treatment regime for chronic diseases using data extracted from an electronic medical record data set during a fixed period of time.

This work was motivated by the National Health and Nutrition Examination Survey (NHANES). The study was designed to assess the health status of adults and children across the United States. Timbie et al. (2010) simulates a cohort of subjects diagnosed with diabetes that mimics the third wave of the NHANES and uses this cohort to evaluate the attainability of the risk factors control using the available treatments. Specifically, Timbie and colleagues’ study was designed to manage the risk factors for vascular complications such as blood pressure, cholesterol and blood glucose (Grundy et al., 2004; Hunt, 2008). In this manuscript, we simulate a cohort similar to Timbie’s. Our focus is on constructing a DTR for lowering hemoglobin A1c among patients with diabetes.

One of the challenges in constructing an optimal regime is to avoid treatments that are optimal in the short term but do not result in an optimal long-term outcome. To address this challenge, Murphy (2003) introduced a method based on backward induction (dynamic programming) to estimate the optimal regime using experimental or observational data.
Murphy’s method starts from the last decision point and finds the treatment option that optimizes the outcome and goes backward in time to find the best treatment regime for all the decision points (Bather 2000, Jordan 2002). More specifically, backward induction maps the covariate history of each individual to an optimal regime. Another method was introduced by Robins (2004) using structural nested mean models (SNMMs). The key notion in Murphy’s method and SNMMs is that the optimal regime can be characterized by just modeling the difference between the outcome under different treatment regimes, rather than the full outcome model. Robins (2004) proposes G-estimation as a tool to estimate the parameters of SNMMs, while Murphy (2003) uses a least square characterization method (Moodie et al., 2007).

Q-learning, a reinforcement learning algorithm, is also widely used in constructing the optimal regime (Murphy et al., 2006; ZHAO et al., 2009; Chakraborty et al., 2010; Nahum-Shani et al., 2012). Q-learning is an extension of the standard regression method that can be used with longitudinal data in which treatments vary over time. Goldberg and Kosorok (2011) introduce a new Q-learning method that can be used when individuals are subject to right censoring. Their new method creates a pseudo population in which everyone has an equal number of decision points and they show that the results obtained by the pseudo population can be translated to the original problem (Zhao et al., 2011; Zhang et al., 2012, 2013, 2014). Schulte et al. (2014) provides a self-contained description of different methods for estimating the optimal treatment regime in finite horizon settings.

The existing methods in the statistics literature are specifically designed to estimate the optimal treatment regime that optimizes a utility function over a fixed period of time. However, in this manuscript our inferential goal is to construct the optimal regime in infinite horizon settings while our data are collected over a fixed period of time. This requires a methodology that estimates the Q-function and the optimal decision rule without the time index. We achieve this by developing an estimating equation that estimates the optimal regime without requiring backward induction from the last to the first decision point. In order to be able to capture the disease dynamic and the long-term treatment effects, our
data should contain a long trajectory of data with many decision points.

The remainder of this manuscript is organized as follows. Section 2 explains the data structure and presents the proposed method. In Section 3, we develop asymptotic properties of our method and describe the gradient descent algorithm that we use to minimize our objective function. We conduct a simulation study in Section 4, where we examine the performance of our method. The last section contains some concluding remarks.

2. CONSTRUCTING THE OPTIMAL REGIME

2.1 Data structure

We study the effect of a time-dependent treatment $A_t$ on a function of outcome. Our data set is composed of $n$ i.i.d trajectories. The $i$th trajectory is composed of the sequence $(X_{i0}, A_{i0}, ..., A_{i(T-1)}, X_{iT})$ where $X_{it}(\cdot)$ is the set of variables measured at the $t$th decision point and $A_{it}$ is the treatment assigned at that decision point after measuring $X_{it}(\cdot)$. $T$ is the maximal number of decision points, and the observed length of trajectories are allowed to be different. At each decision point $t$, we define a variable $S_{it}$ as a summary function of $(X_{i0}, A_{i0}, ..., A_{i(t-1)}, X_{it})$. If a patient dies before the end of the study, say $t$, we set $S_t = \emptyset$ (absorbing state). Given $S_t = s$, $A_t$ takes values in $A_s = \{0, 1, 2, ..., m_s\}$ for all $t$ where $m_s < \infty$, $\forall s \in S$. The treatment and the summary function history through $t$ are denoted by $\bar{A}_t$ and $\bar{S}_t$, respectively. We use small letters to refer to the possible values of the corresponding capital letter random variable. From this point forward, for simplicity of notation, we drop the subscript $i$.

2.2 Potential outcomes

We use a counterfactual or potential outcomes framework to define the causal effect of interest and to state assumptions. Potential outcome models are introduced by Neyman (1990) and Rubin (1978) for time-independent treatment and later extended by Robins (1986, 1987) to assess the time-dependent treatment effect from experimental and observational longitudinal studies.
Associated with each fixed value of the treatment vector $\bar{a}_m$, we conceptualize a vector of the potential outcomes $S_{m+1}(\bar{a}_m) = (S_2(a_1), \ldots, S_{m+1}(\bar{a}_m))$, where $S_{t+1}(\bar{a}_t)$ is the value of the summary function at the $(t + 1)$th decision point that we would have observed had the individual been assigned the treatment history $\bar{a}_t$.

In the potential outcome framework, we make the following assumptions to identify the causal effect of a dynamic regime:

1. **Consistency**: $S_{t+1}(\bar{A}_t) = S_{t+1}$ for each $t$

2. **Sequential randomization**: For each $t$, $A_t$ is conditionally independent of $S_{t+1}(\bar{a}_t)$ given $S_t$.

These assumptions link the potential outcome and the observed data (Robins, 1994, 1997). Assumption 1 means that the potential outcome of a treatment regime corresponds to the actual outcome if assigned to that regime. Assumption 2 means that within levels of $S_t$, treatment at time $t$, $A_t$, is randomized. Throughout this manuscript, we assume that these identifiability assumptions hold.

Besides the above assumptions, we assume that the data generating law satisfies the following assumptions:

**A.1 Markovian assumption**: For each $t$ and $\forall a \in A_{S_t(A_t-1)}$

$$S_{t+1}(a) \perp S_0, A_0, S_1(A_0), \ldots, A_{t-2}, S_{t-1}(A_{t-2}), A_{t-1} \quad \text{given } S_t(A_t-1). \quad (1)$$

We assume that the support of $S_t$ is the same for all $t$s and denote it as $\mathcal{S}$. Assuming that $S_t(A_t-1)$ is rich enough so that the sequential randomization assumption holds, for each $\mathcal{B} \subset \mathcal{S}$ and each $t$, we have

$$p(S_{t+1}(a) \in \mathcal{B}|S_t(A_t-1) = s) = p(S_{t+1}(a) \in \mathcal{B}|S_t(A_t-1) = s, A_t = a)$$

$$= p(S_{t+1} \in \mathcal{B}|S_t(A_t-1) = s, A_t = a),$$

**A.2 Time homogeneity**: For each $s \in \mathcal{S}$ and $a \in A_s$,
\[ p(S_{t+1} \in B | S_t(A_{t-1}) = s, A_t = a) = p(S' \in B | S = s, A = a), \]

where \( S \) and \( S' \) are the summary functions at the previous and the next time, respectively. From this point forward, we refer to \( S \) as a state variable.

**A.3 Positivity assumption:** Let \( p_{A|S}(a|s) \) be the conditional probability of receiving treatment \( a \) given \( S = s \). For each action \( a \in A_s \) and for each possible value \( s \), \( p_{A|S}(a|s) > 0 \).

The last equality in assumption A.2 means that the conditional distribution of the \( S \)s does not depend on \( t \). A.3 ensures that all treatment options in \( A_s \) have been assigned to some patients. It is also known as an exploration assumption since for each \( S = s \), all actions in \( A_s \) are possible that improve our knowledge about different treatments.

### 2.3 The likelihood

Under the Markovian assumption, the distribution of the trajectory \( \{S_t, A_t\}_0^T \) given \((S_{t-1}, A_{t-1})\) is a function of the conditional density of \( S' \) given \( S \) and \( A \), say \( f_{S'|S,A} \), and the density \( p_{A|S}(a|s) \). For a given \( T \), the likelihood of the observed trajectory \( \{s_0, a_0, ..., a_{T-1}, s_T\} \) is given by

\[
    f_S(s_0)p(a_0|s_0) \left( \prod_{k=1}^{T-1} f_{S'|S,A}(s_k | s_{k-1}, a_{k-1})p(a_k | s_k) \right) f_{S'|S,A}(s_T | s_{T-1}, a_{T-1}). \tag{2}
\]

Expectations with respect to this distribution are denoted by \( \mathbb{E} \).

The treatment regime (policy), \( \pi \), is a deterministic decision rule where for every \( s \), the output \( \pi(s) \) is an action \( a \in A_s \), where \( A_s \) is the space of feasible actions (Robins 2004; Schulte et al. 2014). The likelihood of the trajectory \( \{s_0, a_0, ..., a_{T-1}, s_T\} \) corresponding with this law, for a given \( T \), is

\[
    f(s_0)I(a_0 = \pi(s_0)) \left( \prod_{k=1}^{T-1} f_{S'|S,A}(s_k | s_{k-1}, a_{k-1})I(a_k = \pi(s_k)) \right) f_{S'|S,A}(s_T | s_{T-1}, a_{T-1}). \tag{3}
\]

Expectations with respect to this distribution are denoted by \( \mathbb{E}_\pi \). Note the likelihood (3) is not well-defined and it may be identical to zero unless A.3 holds for each possible value \( s \).
and \( a \in A_s \).

### 2.4 Preliminaries

We define the reward value as a known function of \((S_{t-1}, A_{t-1}, S_t)\) at each time \( t \) and denote it by \( R_t = r(S_{t-1}, A_{t-1}, S_t) \). The reward value is a longitudinal outcome that is coded such that high values are preferable. We set \( R_t = 0 \) if \( S_{t-1} = \emptyset \).

The action-value function at time \( t \), \( Q^\pi_t(s,a) \), is defined as an expected value of the cumulative reward if taking treatment \( a \) at state \( s \) at time \( t \) and following the policy \( \pi \) afterward. In other words, \( Q^\pi_t(s,a) \) quantifies the quality of policy \( \pi \) when \( S_t = s \) and \( A_t = a \). Hence, \( Q^\pi_t(s,a) \) is defined as \( \mathbb{E}_\pi \left[ \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} | S_t = s, A_t = a \right] \) where \( \gamma \) is called a discount factor, \( 0 < \gamma < 1 \), which is fixed a priori by the researcher. Note that by definition of the reward function, \( Q^\pi_t(\emptyset,a) = 0 \). Under the Markovian assumption, the action-value function does not depend on \( t \). Thus we can drop the subscript \( t \) and denote it by \( Q^\pi(s,a) \).

Note that the action-value function \( Q^\pi(.,.) \) has a finite value when \( \gamma < 1 \) and the rewards are bounded.

The discount factor \( \gamma \) balances the immediate and long-term effect of treatments on the action-value function. If \( \gamma = 0 \), the objective would be maximizing the immediate reward and ignoring the consequences of the action on future rewards or outcomes. As \( \gamma \) approaches one, future rewards become more important. In other words, \( \gamma \) specifies our inferential goal. For example, suppose there are two treatment regimes \( \pi_1 \) and \( \pi_2 \) where \( \pi_1 \) suggests taking treatments that have high treatment effect and high chance of severe side effects in the future (i.e., small rewards) and \( \pi_2 \) suggests taking treatments that have lower treatment effect but no side effects. In this hypothetical situation, the action-value function for small \( \gamma \)s, say \( \gamma = 0.1 \), has higher value under \( \pi_1 \) than \( \pi_2 \) since we are ignoring the future rewards. However, for larger \( \gamma \)s, say \( \gamma = 0.7 \), the action value function under \( \pi_2 \) may have higher value than \( \pi_1 \). Hence a researcher has to choose \( \gamma \) based on the goal of the study. We have discussed the effect of the choice of \( \gamma \) on the estimated optimal regime in \[ \text{Appendix E} \].
The action-value function can be written as

\[ Q^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{k=1}^\infty \gamma^{k-1} R_{t+k} | S_t = s, A_t = a \right] \]

\[ = \mathbb{E}_\pi \left[ R_{t+1} + \gamma \sum_{k=1}^\infty \gamma^{k-1} R_{t+k+1} | S_t = s, A_t = a \right] \]

\[ = \mathbb{E} \left[ R_{t+1} + \gamma \mathbb{E}_\pi \left\{ \sum_{k=1}^\infty \gamma^{k-1} R_{t+k+1} | S_{t+1}, A_{t+1} = \pi(S_{t+1}) \right\} | S_t = s, A_t = a \right] \]

\[ = \mathbb{E} \left[ R_{t+1} + \gamma Q^\pi(S_{t+1}, \pi(S_{t+1})) | S_t = s, A_t = a \right]. \]

The last equation is known as Bellman equation for \( Q^\pi(s, a) \) (Sutton and Barto, 1998; Si, 2004). The inner expectation quantifies the quality of policy \( \pi \) at time \( t+1 \), in state \( S_{t+1} \) and with treatment \( \pi(S_{t+1}) \). Taking treatment \( \pi(S_{t+1}) \) at time \( t+1 \) ensures treatment policy \( \pi \) is followed in the interval \( (t, t+1] \). Hence, \( Q^\pi(s, a) \) can be interpreted as the expected value of the cumulative rewards when taking treatment \( a \) at state \( s \) at time \( t \) and following the policy \( \pi \) afterward.

Our goal is to construct a treatment policy that, if implemented, would lead to an optimal action-value function for each pair \( (s, a) \). Accordingly, the optimal action-value function can be defined as

\[ Q^*(s, a) = \max_\pi \mathbb{E}_\pi \left[ \sum_{k=1}^\infty \gamma^{k-1} R_{t+k} | S_t = s, A_t = a \right] \]

\[ = \max_\pi \mathbb{E}_\pi \left[ R_{t+1} + \gamma \mathbb{E}_\pi \left\{ \sum_{k=1}^\infty \gamma^{k-1} R_{t+k+1} | S_{t+1}, A_{t+1} = a^* \right\} | S_t = s, A_t = a \right] \]

\[ = \mathbb{E} \left[ R_{t+1} + \gamma \max_\pi \mathbb{E}_\pi \left\{ \sum_{k=1}^\infty \gamma^{k-1} R_{t+k+1} | S_{t+1}, A_{t+1} = a^* \right\} | S_t = s, A_t = a \right] \]

\[ = \mathbb{E} \left[ R_{t+1} + \gamma Q^*(S_{t+1}, a^*) | S_t = s, A_t = a \right], \]

where \( a^* \in \arg \max_a Q^*(S_{t+1}, a) \). Taking treatment \( a^* \) at time \( t+1 \) ensures that we are taking an optimal treatment in the interval \( (t, t+1] \). The last equality follows from the definition of \( Q^*(s, a) \) and can also be written as
\[ Q^*(s, a) = \mathbb{E} \left[ R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') | S_t = s, A_t = a \right]. \tag{4} \]

Note that the only distribution involved in the \( \mathbb{E} \) is \( f_{S'|S,A} \). Denote any policy \( \pi^* \) for which

\[ Q^*(s, a) = \mathbb{E}_{\pi^*} \left[ \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} | S_t = s, A_t = a \right] \]

as an optimal policy. Define the optimal policy as \( \pi^*(s) = \arg \max_a Q^*(s, a) \) and the optimal value function, for state \( s \), as \( V^*(s) = Q^*(s, \pi^*(s)) \).

The action-value function \( Q^*(s, a) \) can be estimated by turning the recurrence relation (4) into an update rule that relies on estimating the conditional density \( f_{S'|S,A} \). However, when the cardinality of \((S, A)\) and the dimension of \( S_t \) are large, estimation of the conditional densities are infeasible. We refer to this method as the classical approach and explain it in Section 4 (Simester et al., 2006; Mannor et al., 2007). One way to overcome this limitation is to use linear function approximation for \( Q^*(s, a) \) which is discussed in the following subsection.

2.5 The proposed estimating equation

The optimal action-value function (4) is unknown and needs to be estimated in order to construct the optimal regime. Suppose the \( Q^*(., .) \) function can be represented using a linear function of parameters \( \theta_0 \),

\[ Q^*(s, a) = \theta_0^\top \varphi(s, a), \]

where \( \theta_0 \) is the parameter vector of \( p \) dimension and \( \varphi(s, a) \) can be any vector of features summarizing the state and treatment pair \((s, a)\) (Sutton et al., 2009a,b; Maei et al., 2010). Features are constructed such that \( \varphi(\emptyset, a) = 0 \). Accordingly, we define the optimal dynamic treatment regime \( \pi^*(s) \) as \( \arg \max_a \theta_0^\top \varphi(s, a) \).

Now we discuss how to estimate the unknown vector of parameters \( \theta_0 \). First, we define an error term and then we construct an estimating equation. In view of the Bellman equation (4), for each \( t \), we have
\[ \mathbb{E} \left[ R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') - Q^*(S_t, A_t) \middle| S_t = s, A_t = a \right] = 0. \] (5)

Thus, the error term, at time (t+1), in the linear setting can be defined as \( \delta_{t+1}(\theta) = R_{t+1} + \gamma \max_{a'} [\theta^\top \varphi(S_{t+1}, a')] - \theta^\top \varphi(S_t, A_t) \) which is known as temporal difference error in computer science literature. In order to account for the influence of the feature function \( \varphi(S, A) \) in the estimation of the \( \theta \)s, we multiply \( \delta(\theta) \) by \( \varphi(S, A) \) and define \( \theta_0 \) as a value of \( \theta \) such that
\[
D(\theta) = \mathbb{E} \left[ \sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right] = 0. \] (6)

The expectation in the above equation depends on the transition density \( f_{S'|S,A} \) and \( p_{A|S} \).

Note that as in (5),
\[
D(\theta_0) = \mathbb{E} \left[ \sum_{t=0}^{T-1} \{ R_{t+1} + \gamma \max_{a} [\theta_0^\top \varphi(S_{t+1}, a)] - \theta_0^\top \varphi(S_t, A_t) \} \varphi(S_t, A_t)^\top \right] \\
= \sum_{t=0}^{T-1} \mathbb{E} \left[ \{ R_{t+1} + \gamma \max_{a} [\theta_0^\top \varphi(S_{t+1}, a)] - \theta_0^\top \varphi(S_t, A_t) \} \varphi(S_t, A_t)^\top \right] \\
= 0,
\]

Hence, given the observed data, an unbiased estimating equation for \( \theta \) can be defined as
\[
\hat{D}(\theta) = \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right] = 0, \] (7)

where \( \mathbb{P}_n \) is the empirical average.

3. CALCULATION

In practice, sometimes there is no \( \hat{\theta} \) that solves the system of equations (7), and sometimes the solution is not unique. One way to deal with this is to take an approach similar to the least square technique and define \( \hat{\theta} \) as a minimizer of an objective function. As in (6), a simple objective function can be defined as
\[
\mathbb{E} \left[ \sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right] \mathbb{E} \left[ \sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right]^\top
\]
The objective function used in this manuscript is the above function weighted by the inverse of the feature covariance matrix and defined as

\[ M(\theta) = D(\theta)W^{-1}D(\theta)\top, \]  

(8)

where \( W = \mathbb{E} \left[ \sum_{t=0}^{T-1} \varphi(S_t, A_t)\varphi(S_t, A_t)\top \right] \). The weight \( W^{-1} \) improves the performance of the proposed stochastic minimization algorithm (Section 3.1). The function \( M(\theta) \) is a generalization of the objective function presented in [Maei et al. (2010)]

The objective function \( M(\theta) \) can be estimated using the observed \((s_t, a_t)\) by

\[ \hat{M}(\theta) = \hat{D}(\theta)\hat{W}^{-1}\hat{D}(\theta)\top, \]  

(9)

where \( \hat{D}(\theta) = \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \delta_{t+1}(\theta)\varphi(S_t, A_t)\top \right] \) and \( \hat{W} = \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \varphi(S_t, A_t)\varphi(S_t, A_t)\top \right] \). Define \( \hat{\theta} \in \arg\min_{\theta} \hat{M}(\theta) \). Then, the estimated optimal dynamic treatment regime is \( \hat{\pi}(s) = \arg\max_a \hat{\theta}\top\varphi(s,a) \). Lemma 1 in [Appendix C] shows that \( \hat{W} \) and \( \hat{D}(\theta) \) are unbiased estimators of \( W \) and \( D(\theta) \), respectively, and the estimator \( \hat{M}(\theta) \) converges in probability to \( M(\theta) \).

The following theorem presents the consistency and asymptotic normality of the estimator \( \hat{\theta} \). This allows investigators to test the significance of variables for use in this sequential decision making problem. The assumptions A.3 – 7 required in this section are listed in [Appendix C]

**Theorem 1** (Consistency and asymptotic normality) For a map \( \theta \to M(\theta) \), defined in (8), under assumptions A.3 – 6, any sequence of estimators \( \hat{\theta} \) with \( \hat{M}(\hat{\theta}) \leq \hat{M}(\theta_0) + o_p(1) \) satisfies the following statements:

I. \( \sqrt{n}(\hat{\theta} - \theta_0) = O_p(1) \).
II. Under A.7, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Gamma^\top \Sigma \Gamma)$, where

$$\Sigma = \mathbb{E} \left[ \left\{ \sum_t \delta_{t+1} \varphi(S_t, A_t)^\top \right\}^\top \left\{ \sum_t \delta_{t+1} \varphi(S_t, A_t)^\top \right\} \right],$$

$$\Gamma = \left[ I - \gamma W^{-1} \mathbb{E} \left( \sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top \right) \right]^\top W + \gamma^2 \mathbb{E} \left( \sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top \right) W^{-1} \mathbb{E} \left( \sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top \right)^\top - 2\gamma \mathbb{E} \left( \sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top \right)^\top$$

where $I$ is an identity matrix and $\delta_{t+1} = [R_{t+1} + \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) - \theta_0^\top \varphi(S_t, A_t)]$.

The asymptotic variance $\Gamma^\top \Sigma \Gamma$ may be estimated consistently by replacing the expectations with expectations with respect to the empirical measure and $\theta_0$ with its estimate $\hat{\theta}$.

The objective function $M(\theta)$ is a non-convex and non-differentiable function of $\theta$, which makes the estimation process complicated. Standard optimization techniques often fail to find the global minimizer of this function. Thus, in the next section, we present an incremental approach, which is a generalization of the iterative stochastic minimization algorithm greedy gradient Q-learning (GGQ) introduced by Maei et al. (2010) as a tool to minimize $M(\theta)$. Hence, from this point forward, we may refer to our proposed method as GGQ.

3.1 The update rule

We base our minimization procedure on the (approximate) gradient descent approach, in which sub-gradients are defined as Fretche sub-gradients of the objective function $M(\theta)$. Maei et al. (2010) introduces this approach and shows that under some conditions, the algorithm converges to the minimizer of the objective function.
The sub-gradient $\partial M(\theta)$ of $M(\theta)$ with respect to $\theta$ is

$$\partial M(\theta) = -E\left[\sum_t \delta_{t+1}(\theta)\varphi(s_t, a_t)^\top\right] + \gamma E\left[\sum_t \varphi(s_{t+1}, \pi^*_{\theta_k}(s_{t+1}))\varphi(s_t, a_t)^\top\right] \varpi,$$

where $\varpi = E[\sum_t \varphi(s_t, a_t)\varphi(s_t, a_t)^\top]^{-1} E[\sum_t \delta_{t+1}(\theta)\varphi(s_t, a_t)^\top]$. Using the weight-doubling trick introduced by Sutton et al. (2009b), we summarize the steps toward minimizing the objective function $M(\theta)$ as follows:

1. Set initial values for the $p$ dimensional vectors of $\theta$ and $\varpi$. Using grid search, the initial value $\theta_1$ can be set as the one that minimizes the objective function and the initial value $w_1 = P_n[\sum_t \varphi(s_t, a_t)\varphi(s_t, a_t)^\top]^{-1} P_n[\sum_t \delta_{t+1}(\theta_0)\varphi(s_t, a_t)^\top]$.

2. Start from the first individual’s trajectory and obtain the $\theta_{k+1}$ from the following iterative equations

$$\theta_{k+1} = \theta_k + \alpha_k \sum_t \left[\delta_{t+1}(\theta_k)\varphi(s_t, a_t) - \gamma \{\varpi_t^\top \varphi(s_t, a_t)\} \varphi(s_{t+1}, \pi_{\theta_k}^* (s_{t+1}))\right]$$  

$$\varpi_{k+1} = \varpi_k + \beta_k \sum_t \left[\delta_{t+1}(\theta_k) - \{\varphi(s_t, a_t)^\top \varpi_t\} \varphi(s_t, a_t)^\top\right],$$

where $\alpha_k$ and $\beta_k$ are tuning parameters (step sizes) and $\pi_{\theta_k}^* (.)$ is the optimal policy estimated as a function of $\theta_k$.

3. Continue updating the parameters using step 2 to the last individual.

4. Continue steps 2 and 3 till convergence.

The tuning parameters (step sizes) $\alpha_k$ and $\beta_k$ need to satisfy the assumptions $P.1$-$P.4$ listed in Appendix A.

3.2 Convergence criteria

Here, we introduce two stopping rules for the update equations (10) and (11). A standard, common convergence criterion is defined based on the Euclidean distance of the parameter
vector of two consecutive iterations: \( \| \theta_k - \theta_{k-1} \|_2 < c_1 \), where \( c_1 \) is a constant. Another criterion can be defined as \( | \text{Norm}_k - \text{Norm}_{k-1} | < c_2 \) where

\[
\text{Norm}_k = \left\| \sum_t \left[ R_{t+1} + \gamma \max_a [\theta_k^\top \varphi(S_{t+1}, a)] - \theta_k^\top \varphi(S_t, A_t) \right] \varphi(S_t, A_t)^\top \right\|_2
\]

and \( c_2 \) is a constant. Note that the empirical average of the vector inside the norm is a Monte Carlo estimate of \( \mathbb{E} \left[ \sum_t \delta_{t+1}(\theta) \varphi(S_t, A_t) \right] \), which represents an error in the updated value of \( \theta \). This vector should converge to zero, so its norm is a measure of the distance from the minimizer of \( M(\theta) \).

4. SIMULATION STUDY

We simulate a cohort of patients with diabetes and focus on constructing a dynamic treatment regime for maintaining the hemoglobin A1c below 7\%. The A1c-lowering treatments that we consider in this manuscript are similar to those of Timbie et al. (2010) and include metformin, sulfonylurea, glitazone and insulin. We used treatment discontinuation rates to measure patients’ intolerance to treatment and reflect both the side effects and burdens of treatment. The treatment efficacies and discontinuation rates are extracted from Kahn et al. (2006) and Timbie et al. (2010). We assume that patients who discontinue a treatment do not drop out but just take the next available treatment. This simulated data mimics the third wave of the NHANES.

4.1 Overview of the simulation.

Our study consists of 20 decision points, and the time between each decision point is 3 months. Eligible individuals start with metformin and augment with treatments sulfonylurea, glitazone and insulin through the follow-up. At each decision point, there are two treatment options: 1) augment the treatment 2) continue the treatment received at the previous decision point. The discontinuation variable \( D \) is generated from a Bernoulli distribution given at the last augmented treatment. \( NAT_t \) is the number of augmented treatments by the end of interval \( t \) where \( NAT_t \in \{0, 1, 2, 3, 4\} \). The variable \( A_t \) is the augmented treat-
ment at time $t$. As soon as a treatment is augmented the variable $NAT$ increases by one, whether or not the treatment will be discontinued.

4.2 Generative model.

Here are the steps we take to generate the dataset:

- **Baseline variables:** Variables ($BP_0, Weight_0, A1c_0$) are generated from multivariate normal distribution with mean $(13,160,9.4)$ and the covariance matrix $\text{diag}(1,1,1)$ where BP is the systolic blood pressure. Also, $NAT_0 = D_0 = 0$.

- **Assigned treatment at time $t$:** Given the state variable $NAT_t$, the sets of available treatments are $A_{NAT_t=0} = \{0, \text{Metformin}\}$, $A_{NAT_t=1} = \{0, \text{Sulfonylurea}\}$, $A_{NAT_t=2} = \{0, \text{Glitazone}\}$, $A_{NAT_t=3} = \{0, \text{Insulin}\}$, and $A_{NAT_t=4} = \{0\}$, where 0 means continue with the treatment received at the previous decision point. Although the ideal $A1c$ level is below 7%, Timbie et al. (2010) raises concern about the feasibility and polypharmacy burden needed for treating patients whose $7 < A1c < 8$. Our simulation study investigates the optimal treatment regime for these patients. More specifically,

  - if $A1c_t < 7$, the treatment is not augmented because $A1c$ is under control and $NAT_t = NAT_{t-1}$.

  - if $A1c_t > 8$ and $NAT_{t-1} < 4$, the treatment is augmented with the next available treatment. Hence, $NAT_t = NAT_{t-1} + 1$. Note that these are patients whose $A1c$ is too high. Thus, the only option is augmenting their treatment.

  - If $7 < A1c_t < 8$ and $NAT_{t-1} < 4$, then a binary variable $Z_t$ is generated from $Z_t \sim Ber\left(\frac{\exp[-0.2A1c_{t-1}+0.5NAT_{t-1}+0.5D_{t-1}]}{1+\exp[-0.2A1c_{t-1}+0.5NAT_{t-1}+0.5D_{t-1}]}, \right)$, where $D_t$ denotes the discontinuation indicator. $Z_t = 1$ implies that the patient continues with the same treatment as time $t-1$ and we set $A_t = 0$ ($NAT_t = NAT_{t-1}$), while $Z_t = 0$ implies that the patient takes the next available treatment (treatment is augmented) and we set $NAT_t = NAT_{t-1} + 1$. For example, if $7 < A1c_t < 8$ and $NAT_{t-1} = 3$, a patient can be assigned to either augmenting the treatment taken at time $t-1$
with $A_t = \text{insulin}$ or continuing with the same treatment as time $t - 1$ ($A_t = 0$), depending on the generated variable $Z_t$.

**Note:** When $Z_t = 1$, no new treatment is added. Hence, $\mathbb{E}[A_{1c_t}|Z_t = 1, A_{1c_{t-1}}] = \mathbb{E}[A_{1c_t}|A_t = 0, A_{1c_{t-1}}] = A_{1c_{t-1}}$.

- **Treatment discontinuation indicator at time $t$:** A binary variable $D_t$ is generated from a Bernoulli distribution given the last augmented treatment. For all $t$, the treatment discontinuation rates are $p(D_t|A_{t-1} = \text{metformin}) = p(D_t|A_{t-1} = \text{sulfonylurea}) = p(D_t|A_{t-1} = \text{glitazone}) = 0.20$, and $p(D_t|A_{t-1} = \text{insulin}) = 0.35$.

  **Note:** We assume no treatment discontinuation for patients who at time $t$ are taking the same treatment as $t - 1$ (i.e., $p[D_t = 1|A_{t-1} = 0] = 0$).

- **Intermediate outcome $A_{1c}$ at time $t$:** To avoid variance inflation through time, we use the following generative model for $A_{1c}$ at time $t$, $A_{1c_t} = \frac{A_{1c_{t-1}} - \mu_{t-1} + \epsilon}{\sqrt{1 + \sigma^2}} + \mu_t$, where $\epsilon \sim N(0, \sigma^2 = 0.5)$ and

$$
\mu_t = \mathbb{E}[A_{1c_t}|A_{1c_{t-1}}, NAT_{t-1}, A_t, D_t] = \begin{cases} 
\mu_{t-1}(1 - \tau_{A_t}) & \text{if } A_{1c_{t-1}} > 7, NAT_{t-1} < 4, A_t \neq 0, D_t \neq 1 , \\
\mu_{t-1} & \text{o.w.}
\end{cases}
$$

with $\tau_{A_t}$ being the treatment effect of the augmented treatment $A_t$. The treatment effects of metformin, sulfonylurea, glitazone* and insulin are 0.14, 0.20, 0.02, and 0.14, respectively. Note that the treatment effects are reported as a percentage reduction in $A_{1c}$. The treatment effect of glitazone is listed as 0.12 in Timbie et al. (2010), which is similar to metformin. However, in order to study the effect of the treatment discontinuation on the optimal regime, we set its treatment effect to 0.02 and, from now on, denote it by glitazone*.

- **Time-varying variables at time $t$:** $BP_t = (BP_{t-1} + \epsilon)/(\sqrt{1 + \sigma^2})$ and $Weight_t = (Weight_{t-1} + \epsilon)/(\sqrt{1 + \sigma^2})$.

- **Reward function at time $t$:** In order to be able to find an optimal treatment regime, we need an operational definition of controlled $A_{1c}$. Hence, we define the following
reward function at time $t$ as a function of $A1c_t$ and $D_t$,

$$R_t = 1 \text{ if } A1c_t < 7, \quad -5 \text{ if } 7 < A1c_t \& D_t = 1 \text{ and zero otherwise.}$$

This reward structure helps us to identify treatments whose discontinuation rate outweighs their efficacy.

Note that the state space at time $t$ includes $S_t = (NAT_t, D_t, A1c_t, BP_t, Weight_t)$. However, the Markov property holds with $(NAT_t, D_t, A1c_t)$, and variables $BP$ and $Weight$ are noise variables.

We generate two datasets of sizes 2,000 and 10,000 and compare the quality of the estimated optimal treatment policy using the proposed GGQ and the classical approach. The latter, also known as action-value iteration method, turns the recurrence relation of (4) into an update rule as

$$Q_{k+1}^*(s, a) = \mathbb{E} \left[ r(s, a, S') + \gamma \max_{a'} Q_k^*(S', a') | S = s, A = a \right]$$

$$= \sum_{s'} P_{S'|S,A}(s'|s, a) \left[ r(s, a, S') + \gamma \max_{a'} Q_k^*(s', a') \right],$$

where $r()$ is the reward function. This procedure can be summarized as

1. set $Q_1^*(s, a) = 0$ for all $(s, a) \in (S, A_s)$
2. for each $(s, a) \in (S, A_s)$, $q \leftarrow Q_k^*(s, a)$
3. $Q_{k+1}^*(s, a) \leftarrow \sum_{s'} P_{S'|S,A}(s'|s, a) \left[ r(s, a, S') + \gamma \max_{a'} Q_k^*(s', a') \right]$
4. repeat 2 and 3 until $\max_{s,a} |Q_{k+1}^*(s, a) - q| < \epsilon$ where $\epsilon$ is a small positive value
5. for each $s$, $\hat{\pi}(s) = \arg \max_a Q_{k+1}^*(s, a)$.

The above 5-step procedure is similar to the one presented in Chapter 4 of Sutton and Barto (1998). Note that the classical approach requires estimation of the transition probabilities $P_{S'|S,A}$, which limits its usage to cases where state and action space is not too large. We categorize the continuous variables ($BP, Weight, A1c$) and estimate $P_{S'|S,A}$ nonparametrically.
Figure 1: Simulation: Estimated optimal treatment (op.txt) for different states. The shaded bar represents the evidence in the simulated data for each of the treatment choices as labeled in the legend. The upper and lower horizontal axes are the discontinuation indicator and the categorized A1c (Cat.A1c), respectively. The vertical axes on the right and left hand side give NAT and the percentage of time that the treatment choices are selected as the optimal choice.

The variables ($BP, Weight$) are categorized based on the percentiles (30, 80) and denoted as $(Cat.BP, Cat.Weight)$. The categorized $A1c$ (Cat.$A1c$) is formed by breaking the variable $A1c$ on $(-\infty, 7, 7.2, 7.5, 7.7, 7.8, 9, +\infty)$. Hence the state variable used in the classical approach is $S_t^{Class} = (NAT_t, D_t, Cat.BP, Cat.Weight, Cat.A1c)$. Note that $Cat.A1c \in \{2, 3, 4, 5\}$ corresponded to $7 < A1c < 8$.

Unlike the classical approach, the optimal treatment policy using our proposed GGQ method utilizes the continuous state variable $S_t = (NAT_t, D_t, A1ct, BP_t, Weight_t)$. In our example, we parametrize the optimal action value function $Q^*(s, a)$ using a 72-dimensional vector of parameters and construct the features $\phi(s, a)$ using radial basis functions (Gaussian kernels). See Appendix B for more details. To specify the step sizes of the stochastic minimization algorithm, first we select two functions that satisfy the conditions $P.1-4$ listed in Appendix A and multiply them by $v \in (0, 1)$. Then we run the algorithm for different
Figure 2: Simulation: Monte Carlo approximation of the difference between value functions. $V^{\pi_{\text{GGQ}}}$ and $V^{\pi_{\text{Class}}}$ are the value functions corresponding to the classical and GGQ approaches. The vertical axis represents the triplets of states in the order of $(\text{NAT}, D, \text{Cat. A1c})$.

values of $\nu$ and select the one that minimizes the objective function. In this simulation study we set the step sizes $\alpha_k = \nu / k$ and $\beta_k = \nu / k^{3/4}$ where $\nu$ is set to 0.03.

**True optimal policy.** As the sample sizes increases, the optimal action-value function estimated using the classical approach converges to the true optimal action-value function. Hence, for the purpose of finding the true optimal policy, we generate a large dataset of size 500,000 and estimate transition probabilities $P_{S'|S,A}$ using a nonparametric approach where $S$ is the oracle state $(\text{NAT}_i, D_i, \text{Cat. A1c}_i)$. Then by the 5-step procedure (classical approach), the true optimal policy is approximated and set as our benchmark.

Figure 1 depicts the true and estimated optimal treatment for each discretized oracle state $(\text{NAT}_i, D_i, \text{Cat. A1c}_i)$ using the GGQ and classical approach as in this example, we set the discount factor $\gamma$ to 0.6. Note that the estimated optimal policy using classical and GGQ methods are based on the states $S_i^{\text{Class}}$ and $S_i$, respectively. However, in Figure 1, we averaged it over the noise variables $(BP, Weight)$ and, for comparability, we report the results on the discretized oracle state. The vertical axes on the left hand side is the percentage of time that the treatment choices are selected as optimal. The left vertical,
the upper and lower horizontal axes represent the elements of the state \((NAT, D, Cat.A1c)\), respectively. This plot shows that the proposed GGQ method outperforms the classical approach for moderate sample sizes.

More specifically, the estimated optimal policy using GGQ \((\hat{\pi}_{GGQ})\) recommends not augmenting the third \((glitazone^*)\) and fourth \((insulin)\) treatments (the left plots). This makes sense since \(glitazone^*\) has a small treatment effect \((0.02)\) and \(insulin\) has high discontinuation rate \((0.35)\) that out-weighs its efficacy. However, the estimated optimal policy using a classical approach \((\hat{\pi}_{Class})\) recommends augmenting these treatments by some positive probabilities. Specifically, \(\hat{\pi}_{Class}\) augments \(insulin\) about 50% of the time when \(D = 1\) and \(Cat.A1c \in \{2, 3, 4\}\).

Figure 2 presents the difference between the values of the estimated optimal policies \(\hat{\pi}_{GGQ}\) and \(\hat{\pi}_{Class}\). The values are calculated using the Monte Carlo method, where the value of a treatment policy \(\pi\), for each state \(s\), is defined as \(V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}s_t = s, A_t = \pi(s) \right]\). For both sample sizes and all of the states, the value of \(\hat{\pi}_{GGQ}\) is higher than the value of \(\hat{\pi}_{Class}\). This indicates that the estimated optimal policy \(\hat{\pi}_{GGQ}\) has better quality. Figure 3 shows the
Figure 4: Simulation: The confidence intervals of the difference between the estimated optimal action-value function when the treatment is augmented and continued (i.e., \( \hat{Q}^*(s, \text{augment}) - \hat{Q}^*(s, 0) \)). The number on each confidence interval represents the coverage of that interval. The vertical axis represents the triplets of states in the order of \((\text{NAT}, D, \text{Cat. A1c})\).

Our simulation result is consistent with Timbie et al. (2010), and suggests that we should not always augment the treatment when \(7 < A1c < 8\). In other words, depending on the treatment already taken, sometimes we should consider not augmenting the treatment to
avoid side effects.

5. DISCUSSION

We have proposed a new method that can be used to form optimal dynamic treatment regimes in infinite horizon settings (i.e., there is no a priori fixed end of follow up), while our data were collected over a fixed period of time with many decision points. We have assumed that the value of the optimal regime can be presented using a linear function of parameters and we developed an estimating procedure based on temporal difference residuals to estimate the parameters of this function. We developed the asymptotic properties of the estimated parameters and evaluated the proposed method using simulation studies.

This work raises a number of interesting issues. We have derived the asymptotic distribution of the parameters under some assumptions. One important practical problem is to provide a valid inference when the optimal treatment is not unique for some states, (i.e., assumption A.7 is violated). This may lead to non-regular estimators and inflate the Type-I error rate (Bickel et al., 1993). Among others, Robins (2004) and Laber et al. (2010) proposed solutions to this issue. However, the existing methods may not be directly applied to our method and require major modifications. The second issue is how to construct the feature functions. In this manuscript, we used the radial basis functions (Moody and Darken, 1989; Poggio and Girosi, 1990). One simple method is to try different feature functions \( \phi \) and select the one that minimizes the function \( f(\phi) = \min_\theta M(\theta) \) (Parr et al., 2008). Alternatively, one may use support vector regression to approximate the action-value function (Vapnik et al., 1997; Tang and Kosorok, 2012).

The proposed method can be used in settings where the time between decision points is fixed, say 3 months. This assumption often holds (approximately) for some chronic diseases such as diabetes, cyclic fibrosis and asthma. It would, however, be of interest to extend the method to cases with a random decision point (clinic visits) process. Usually, the random time between decision points happens either when doctors decide to schedule the next visit sooner or later than the prespecified time or when patients request an appointment due to,
for example, side effects or acute symptoms. The former is easier to deal with because we have the covariates required to model the visit process. The latter, however, is more difficult and results in non-ignorable missing data because we do not have information about those patients who did not show up. [Robins et al. (2008)] discusses the issue of the random visit process in detail.

ACKNOWLEDGEMENT

This project was partially supported by award number P50 DA010075, from the National Institute on Drug Abuse. The content is solely responsibility of the author and does not necessarily represent the official views of the NIDA or the National Institutes of Health. The author is grateful to Professor Susan Murphy for enlightening discussion.

Appendix A. TUNING PARAMETERS

The tuning parameters $\alpha_k$ and $\beta_k$ in the GGQ algorithm need to satisfy the following assumptions:

P.1 $\alpha_k, \beta_k \forall k$ and are deterministic.

P.2 $\sum_{k=0}^{\infty} \alpha_k = \sum_{k=0}^{\infty} \beta_k = \infty$.

P.3 $\sum_{k=0}^{\infty} (\alpha_k^2 + \beta_k^2) < \infty$.

P.4 $\alpha_k / \beta_k \rightarrow 0$.

Appendix B. FEATURE FUNCTIONS

The feature functions are constructed using the radial basis functions and

$$\varphi(s, a) = I(s \neq \emptyset)(\varphi_1(s, a), \varphi_2(s, a), \varphi_3(s, a), \varphi_4(s, a), \varphi_5(s, a), \varphi_6(s, a), \varphi_7(s, a), \varphi_8(s, a), \varphi_9(s, a)),$$
increase the variance of the estimators. In our simulation, we set the estimated parameters. Similarly, decreasing the value of such that increasing the number of quintiles decreases the bias but increases the variance of our estimator.

\[ \varphi_1(s, a) = I(A = 0, NAT = 0)(1, \exp[-h(A1c - q_{11})^2], \exp[-h(A1c - q_{12})^2], \phi(BP), \phi(Weight)) \]
\[ \varphi_2(s, a) = I(A = 0, NAT = 1)(1, \exp[-h(A1c - q_{21})^2], \exp[-h(A1c - q_{22})^2], \exp[-h(A1c - q_{23})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_3(s, a) = I(A = 0, NAT = 2)(1, \exp[-h(A1c - q_{31})^2], \exp[-h(A1c - q_{32})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_4(s, a) = I(A = 0, NAT = 3)(1, \exp[-h(A1c - q_{41})^2], \exp[-h(A1c - q_{42})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_5(s, a) = I(A = 0, NAT = 4)(1, \exp[-h(A1c - q_{51})^2], \exp[-h(A1c - q_{52})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_6(s, a) = I(A = 1, NAT = 0)(1, \exp[-h(A1c - q_{61})^2], \exp[-h(A1c - q_{62})^2], \phi(BP), \phi(Weight)) \]
\[ \varphi_7(s, a) = I(A = 2, NAT = 1)(1, \exp[-h(A1c - q_{71})^2], \exp[-h(A1c - q_{72})^2], \exp[-h(A1c - q_{73})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_8(s, a) = I(A = 3, NAT = 0)(1, \exp[-h(A1c - q_{81})^2], \exp[-h(A1c - q_{82})^2], d, \phi(BP), \phi(Weight)) \]
\[ \varphi_9(s, a) = I(A = 4, NAT = 2)(1, \exp[-h(A1c - q_{91})^2], \exp[-h(A1c - q_{92})^2], d, \phi(BP), \phi(Weight)), \]

where \( \phi(BP) = (\exp[-h(BP - q_{01})^2], \exp[-h(BP - q_{03})^2]) \) and \( \phi(Weight) = (\exp[-h(Weight - q_{w1})^2], \exp[-h(Weight - q_{w3})^2]) \). \( h \) is a positive constant and \( q_{j} \) is the observed \( j \)th quantile of the corresponding variable. For example, \( q_{11} \) and \( q_{12} \) are the first and second quantiles of \( A1c \) given \( A = 0 \) and \( NAT = 0 \). Similarly, \( q_{61} \) and \( q_{63} \) are the first and third quantiles of \( BP \). Note that, in our generative model, \( BP \) and \( Weight \) are independent of \( A \) and \( NAT \).

**Remark 1.** Number of quintiles used in each \( \varphi_k \) and \( \phi(\cdot) \) is a bias-variance trade-off such that increasing the number of quintiles decreases the bias but increases the variance of the estimated parameters. Similarly, decreasing the value of \( h \) may decrease the bias but increase the variance of the estimators. In our simulation, we set \( h = 0.5 \).

**Appendix C. ASSUMPTIONS AND LEMMA 1 & 2**

Besides assumptions A.1-2, the following assumptions are required for large sample properties of our estimator.

**A.3** \( \theta_0^\top \varphi(\cdot, \cdot) \) is the optimal Q-function.
A.4 \[ \mathbb{E} \left[ \sum_{t=0}^{T-1} \| \varphi(S_t, A_t) \| \| \varphi(S_t, a) \| \right] < \infty, \text{ for any } a \in A. \]

A.5 The matrix $W$ is of full rank.

A.6 \[ \mathbb{E} \left[ \sum_{t=0}^{T-1} \{ \gamma I_{|\pi^*(S_{t+1})|} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t) \} \varphi(S_t, A_t)^\top \right] \text{ is of full rank.} \]

A.7 The optimal treatment is unique at each decision point.

**Lemma 1** Under assumption A.5 and for each $\theta \in \mathbb{R}^d$, $\hat{M}(\theta) \rightarrow M(\theta)$ in probability as $n \rightarrow \infty$.

**Proof** We show that $\hat{W}$ and $\hat{D}(\theta)$ are unbiased estimators. For trajectory $i$, we have

\[
\mathbb{E}[\hat{D}(\theta)] = \mathbb{E} \left[ \sum_{t=0}^{T-1} \{ R_{t+1} + \max_a [\theta^\top S_{t+1}, a] - \theta^\top \varphi(S_t, A_t) \} \varphi(S_t, A_t)^\top \right] \\
= \sum_{t=0}^{T-1} \mathbb{E} \left[ \{ R_{t+1} + \max_a [\theta^\top S_{t+1}, a] - \theta^\top \varphi(S_t, A_t) \} \varphi(S_t, A_t)^\top \right] \\
= D(\theta)
\]

and

\[
\mathbb{E}[\hat{W}] = \mathbb{E} \left[ \sum_{t=0}^{T-1} \varphi(S_t, A_t) \varphi(S_t, A_t)^\top \right] \\
= \sum_{t=0}^{T-1} \mathbb{E} \left[ \varphi(S_t, A_t) \varphi(S_t, A_t)^\top \right] \\
= W,
\]

Using the law of large numbers, $\hat{W} \xrightarrow{P} W$ and $\hat{D}(\theta) \xrightarrow{P} D(\theta)$. Assuming $W$ is a positive definite matrix, we have $\hat{W}^{-1} \xrightarrow{P} W^{-1}$. Thus $\hat{D}(\theta)\hat{W}^{-1}\hat{D}(\theta)^\top \xrightarrow{P} M(\theta)$, for each $\theta \in \mathbb{R}^d$.

**Lemma 2** Let $\{a_i\}_1^K$ and $\{b_i\}_1^K$ be two sets of elements, then

I. \[ \lim_{\|b\| \rightarrow 0} \frac{\max_{1 \leq i \leq K} |a_i + b_i| - \max_{1 \leq i \leq K} |a_i|}{\|b\|} = 0; \]
II. \(\max_{1 \leq i \leq K} [a_i + b_i] - \max_{i \in \pi^*} [a_i + b_i]\) is non-negative and bounded above by

\[
\max_{i \notin \pi^*} b_i - \max_{i \in \pi^*} b_i \leq \max_{1 \leq i \leq K} b_i - \max_{i \in \pi^*} b_i \leq \max_{1 \leq i \leq K} |b_i|,
\]

where \(\pi^* = \arg \max_{1 \leq i \leq K} a_i\).

**Proof** Part I. Since the set \(\pi^*\) is a subset of \(1 \leq i \leq K\), we have

\[
0 \leq \frac{\max_{1 \leq i \leq K} [a_i + b_i] - \max_{i \in \pi^*} [a_i + b_i]}{\|b\|} = \max_{1 \leq i \leq K} \left[ a_i - a_{i^*} + b_i - \max_{i \in \pi^*} b_i \right] \frac{1}{\|b\|}, \quad \forall i^* \in \pi^*
\]

\[
= \max_{i \in \pi^*} \left[ \max\{b_i - \max_{i \in \pi^*} b_i, \max\{a_i - a_{i^*} + b_i - \max_{i \in \pi^*} b_i\}\} \right] \frac{1}{\|b\|}
\]

\[
= \max_{i \notin \pi^*} \left[ 0, \max\{a_i - a_{i^*}, b_i - \max_{i \in \pi^*} b_i\} \right] \frac{1}{\|b\|}
\]

\[
\leq \max_{i \notin \pi^*} \left[ 0, \max\{a_i - a_{i^*}, \max_{1 \leq i \leq K} b_i\} \right] \frac{1}{\|b\|}
\]

Since \(\frac{a_i - a_{i^*}}{\|b\|} \rightarrow -\infty\) as \(\|b\| \rightarrow 0\) and \(\frac{\max_{i \in \pi^*} b_i}{\|b\|} \leq 1\), part I is proved. Part II can be proved similarly. \(\Box\)

**Appendix D. PROOF OF THEOREM 1**

**Proof** We first show that the objective function \(M(\theta)\) is continuous. Then using the results of Lemma 1 & 2, we show that \(M(\theta)\) satisfies the two required conditions of Theorem 3.2.1 Van Der Vaart and Wellner (1996) which completes the proof of consistency. To prove the asymptotic normality, under the additional assumption A.7, we define a function \(V(b)\) such that \(\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \arg \min_b V(b)\) where \(\arg \min_b V(b)\) is Normally distributed.

First, we show that the function \(M(\theta)\) is continuous by proving the continuity of \(D(\theta)\)
around $\theta = \theta_0$. Since

$$
\|D(\theta) - D(\theta_0)\| = \left\| \sum_{t=0}^{T-1} \mathbb{E} \left[ \{ \max_a [\theta^T \varphi(S_{t+1}, a)] - \max_a [\theta_0^T \varphi(S_{t+1}, a)] 
\right.
\right.
$$

$$
\left. - \theta^T \varphi(S_t, A_t) + \theta_0^T \varphi(S_t, A_t) \} \varphi(S_t, A_t)^T \right]\right\|
$$

by the Cauchy-Schwartz inequality and the fact that $| \max_a f(a) - \max_a g(a) | \leq \max_a | f(a) - g(a) |$, we have

$$
\mathbb{E} \left[ \max_a [\theta^T \varphi(S_{t+1}, a)] - \max_a [\theta_0^T \varphi(S_{t+1}, a)] \right] \leq \| \theta - \theta_0 \| \mathbb{E} \left[ \sum_a \| \varphi(S_{t+1}, a) \| \right],
$$

which under assumption A.4 implies the continuity of $D(\theta)$ around $\theta = \theta_0$.

**Part I (Consistency).** We show that the $M(\theta)$ satisfies the two conditions listed in Theorem 3.2.1 Van Der Vaart and Wellner (1996). For the **first** condition, we need to show that for some $\epsilon > 0$ and $c > 0$ with $\| \theta - \theta_0 \| < \epsilon$,

$$
D(\theta_0)W^{-1}D(\theta_0)^T - D(\theta)W^{-1}D(\theta)^T \leq -c\| \theta - \theta_0 \|^2,
$$

and since $D(\theta_0) = 0$,

$$
D(\theta)W^{-1}D(\theta)^T \geq c\| \theta - \theta_0 \|^2.
$$

(Appendix D.1)

The LHS of the above inequality can be written as

$$
D(\theta)W^{-1}D(\theta)^T = [D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]W^{-1}[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]^T
$$

$$
- \dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}D(\theta_0 - \theta_0)^T + 2\dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]^T
$$

where

$$
\dot{D}_{\theta_0}(b) = \mathbb{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma \max_{a \in \pi^*(S_{t+1})} b^T \varphi(S_{t+1}, a) - b^T \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^T \right],
$$

27
and \( \pi^* (S_{t+1}) = \arg \max_a \theta_0^T \varphi (S_{t+1}, a) \). Note that \( \pi^* (S_{t+1}) \) may be a set of actions. Now, we show that \( \| D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0) \| = o(\| \theta - \theta_0 \|) \).

\[
\frac{\| D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0) \|}{\| \theta - \theta_0 \|} = \gamma \mathbb{E} \left[ \sum_{t=0}^{T-1} \left\{ \max_a (\theta_0 + b \| \theta - \theta_0 \|)^T \varphi (S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} \theta_0^T \varphi (S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} b^T \varphi (S_{t+1}, a) \| \theta - \theta_0 \| \right\} \varphi (S_t, A_t)^T \right],
\]

where \( b = (\theta - \theta_0) / \| \theta - \theta_0 \| \). Then, since \( \forall a, a' \in \pi^*(S_{t+1}) \), we have \( \theta_0^T \varphi (S_{t+1}, a) = \theta_0^T \varphi (S_{t+1}, a') \), the following equality holds

\[
\max_{a \in \pi^*(S_{t+1})} \theta_0^T \varphi (S_{t+1}, a) + \max_{a \in \pi^*(S_{t+1})} (\theta - \theta_0)^T \varphi (S_{t+1}, a) = \max_{a \in \pi^*(S_{t+1})} \theta^T \varphi (S_{t+1}, a).
\]

Thus,

\[
\frac{\| D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0) \|}{\| \theta - \theta_0 \|} = \gamma \mathbb{E} \left[ \sum_{t=0}^{T-1} \left\{ \max_a (\theta_0 + b \| \theta - \theta_0 \|)^T \varphi (S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} \theta_0^T \varphi (S_{t+1}, a) \right\} \varphi (S_t, A_t)^T \right]
\]

\[
\leq \gamma \mathbb{E} \left[ \sum_{t=0}^{T-1} \left\{ \max_a \left( \frac{\theta_0}{\| \theta - \theta_0 \|} + b \right)^T \varphi (S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} \left( \frac{\theta_0}{\| \theta - \theta_0 \|} + b \right)^T \varphi (S_{t+1}, a) \right\} \varphi (S_t, A_t)^T \right].
\]

Using Lemma 2, we have

\[
\lim_{\| \theta - \theta_0 \| \to 0} \left[ \max_a \left( \frac{\theta_0}{\| \theta - \theta_0 \|} + b \right)^T \varphi (S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} \left( \frac{\theta_0}{\| \theta - \theta_0 \|} + b \right)^T \varphi (S_{t+1}, a) \right] = 0.
\]

We just showed that \( \| D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0) \| = o(\| \theta - \theta_0 \|) \) and since \( W^{-1} \) is of full rank matrix, it results

\[
[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0)] W^{-1} [D(\theta) - D(\theta_0) - \dot{D}_{\theta_0} (\theta - \theta_0)]^T = o(\| \theta - \theta_0 \|^2).
\]
Now, we need to show that \( \dot{D}_{\theta_0}(\theta - \theta_0) \dot{D}_{\theta_0}(\theta - \theta_0)^\top \geq c''||\theta - \theta_0||^2 \). By definition,

\[
\dot{D}_{\theta_0}(\theta - \theta_0) = \mathbf{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma \max_{a \in \pi^*(S_{t+1})} (\theta - \theta_0)^\top \varphi(S_{t+1}, a) - (\theta - \theta_0)^\top \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right]
\]

\[
= (\theta - \theta_0)^\top \mathbf{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|=1} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right]
\]

\[
+ \gamma \mathbf{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|>1} \max_{a \in \pi^*(S_{t+1})} \frac{(\theta - \theta_0)^\top}{||\theta - \theta_0||} \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right] ||\theta - \theta_0||.
\]

Let

\[
M_1 = \mathbf{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|=1} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right],
\]

\[
M_2 = \mathbf{E} \left[ \sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|>1} \max_{a \in \pi^*(S_{t+1})} \frac{(\theta - \theta_0)^\top}{||\theta - \theta_0||} \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right].
\]

Then

\[
\dot{D}_{\theta_0}(\theta - \theta_0) \dot{D}_{\theta_0}(\theta - \theta_0)^\top = (\theta - \theta_0)^\top M_1 M_1^\top (\theta - \theta_0) + 2\gamma (\theta - \theta_0)^\top M_1 M_2^\top ||\theta - \theta_0|| + \gamma^2 ||\theta - \theta_0||^2 M_2 M_2^\top.
\]

Assuming that \( M_1 \) is of full rank (Assumption A.6), we have

\[
(\theta - \theta_0)^\top M_1 M_1^\top (\theta - \theta_0) \geq ||\theta - \theta_0||^2 \lambda_{\text{min}},
\]

where \( \lambda_{\text{min}} \) is the smallest eigenvalue of \( M_1^\top M_1 \). Also, using the singular value decomposition we have

\[
(\theta - \theta_0)^\top M_1 M_2^\top ||\theta - \theta_0|| \leq ||\theta - \theta_0||^2 \sqrt{\lambda_{\text{max}} ||M_2||},
\]

where \( \lambda_{\text{max}} \) is a maximum eigenvalue of \( M_1 M_1^\top \). Thus

\[
\dot{D}_{\theta_0}(\theta - \theta_0) \dot{D}_{\theta_0}(\theta - \theta_0)^\top \geq ||\theta - \theta_0||^2 \left[ \lambda_{\text{min}} - 2\gamma \sqrt{\lambda_{\text{max}} ||M_2||} - \gamma^2 ||M_2||^2 \right].
\]

Therefore, the function \( M(.) \) satisfies the first condition of Theorem 3.2.5 Van Der Vaart.
and Wellner (1996) for any small enough $\gamma$ such that $[\lambda_{\min} - 2\gamma\sqrt{\lambda_{\max}}\|M_2\| - \gamma^2\|M_2\|^2] > 0$. Note that under assumption A.7 when $|\pi^*(S_{t+1})| = 1$, the latter condition is satisfied automatically. Since $W$ is of full rank (Assumption A.5), $\dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}\dot{D}_{\theta_0}(\theta - \theta_0)^\top \geq c\|\theta - \theta_0\|^2$ for $c > 0$.

For the **second** condition, we need to show that for every large enough $n$, sufficiently small $\delta_n$ and $c > 0$

$$\mathbb{E} \sup_{\|\theta - \theta_0\| \leq \delta_n} \left| [\hat{M}(\theta) - M(\theta)] - [\hat{M}(\theta_0) - M(\theta_0)] \right| \leq c\delta_n^2.$$ 

Since by definition $\dot{D}(\hat{\theta}) = D(\theta_0) = 0$, we have

$$\left| [\hat{M}(\theta) - M(\theta)] - [\hat{M}(\theta_0) - M(\theta_0)] \right| = \left| (\hat{D}(\theta) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta) - \hat{D}(\hat{\theta}))^\top - (\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))^\top - (D(\theta) - D(\theta_0))\hat{W}^{-1}(D(\theta) - D(\theta_0))^\top \right|.$$ 

We show that for every large $n$ such that $\|\theta - \hat{\theta}\| \leq \delta_n$ and $\|\theta_0 - \hat{\theta}\| \leq \delta_n$

$$\mathbb{E} \sup_{\|\theta - \hat{\theta}\| \leq \delta_n} [(\hat{D}(\theta) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta) - \hat{D}(\hat{\theta}))^\top] \leq c_1\delta_n^2$$

$$\mathbb{E} \sup_{\|\theta - \theta_0\| \leq \delta_n} [(\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))^\top] \leq c_2\delta_n^2$$

$$\mathbb{E} \sup_{\|\theta - \theta_0\| \leq \delta_n} [(D(\theta) - D(\theta_0))^\top\hat{W}^{-1}(D(\theta) - D(\theta_0))^\top] \leq c_3\delta_n^2$$

where $c_1$, $c_2$ and $c_3$ are positive constants. Here, we show the first inequality and the rest can be shown similarly. By the Cauchy-Schwartz inequality and the fact that $|\max_a f(a) -$
\[ \max_a g(a) \leq \max_a |f(a) - g(a)|, \text{ we have} \]

\[
|\hat{D}(\theta) - \hat{D}(\hat{\theta})| = \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \left\{ \max_a \theta_0^T \varphi(S_{t+1}, a) - \max_a \hat{\theta}_0^T \varphi(S_{t+1}, a) - (\theta - \hat{\theta})^T \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^T \right] \\
\leq \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \left\| \varphi(S_t, A_t) \right\| \left\{ \sum_a \left\| \varphi(S_{t+1}, a) \right\| \|\theta - \hat{\theta}\| + \left\| \varphi(S_t, A_t) \right\| \|\theta - \hat{\theta}\| \right\} \right]
\]

Thus,

\[
|\hat{D}(\theta) - \hat{D}(\hat{\theta})| \leq m(S, A) \|\theta - \hat{\theta}\|,
\]

where

\[ m(S, A) = \mathbb{P}_n \left[ \sum_{t=0}^{T-1} \left\| \varphi(S_t, A_t) \right\| \left\{ \sum_a \left\| \varphi(S_{t+1}, a) \right\| + \left\| \varphi(S_t, A_t) \right\| \right\} \right]. \]

Therefore

\[
\mathbb{E} \sup_{\|\theta - \hat{\theta}\| \leq \delta_n} \left[ (\hat{D}(\theta) - \hat{D}(\hat{\theta}))^T \hat{W}^{-1}(\hat{D}(\theta) - \hat{D}(\hat{\theta})) \right] \leq c_1 \sigma_n^2,
\]

where \( c_1 = \mathbb{E}[m(S, A)^2 \|\hat{W}^{-1}\|]. \) Define \( c = c_1 + c_2 + c_3. \) This shows that we satisfy the second condition of Theorem 3.2.5 Van Der Vaart and Wellner (1996) as well. This completes the proof of consistency.

**Part II (Asymptotic Normality).** Define

\[
\hat{V}(b) = n\hat{D}(\theta_0 + b/\sqrt{n})\hat{W}^{-1}\hat{D}(\theta_0 + b/\sqrt{n})^T - n\hat{D}(\theta_0)\hat{W}^{-1}\hat{D}(\theta_0)^T.
\]

Then,

\[
\hat{D}(\theta_0 + b/\sqrt{n}) = \mathbb{P}_n \left[ \sum_t \left\{ R_{t+1} + \gamma \max_a \theta_0^T \varphi(S_{t+1}, a) - \theta_0^T \varphi(S_t, A_t) - b^T / \sqrt{n} \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^T \right] \\
= \mathbb{P}_n \left[ \sum_t \left\{ \delta_{t+1} - b^T / \sqrt{n} \varphi(S_t, A_t) - \gamma \max_a \theta_0^T \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^T \right]
\]
Using the above equation and the defined \( \delta_{t+1} \), the function \( \tilde{V}(b) \) can be written as

\[
\frac{1}{n\rho_n} \left[ \sum_t \left\{ 2\delta_{t+1} - b^T \sqrt{n} \varphi(S_t, A_t) - \gamma \max_a \theta_0^T \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^T \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \\
\times \frac{1}{n\rho_n} \left[ \sum_t \left\{ -b^T \sqrt{n} \varphi(S_t, A_t) - \gamma \max_a \theta_0^T \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^T \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^T \right]^T 
\]

The \( \tilde{V}(b) \) can be decomposed to the following parts:

I. \( n\rho_n \left[ \sum_t b^T \sqrt{n} \varphi(S_t, A_t) \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \frac{1}{n\rho_n} \left[ \sum_t b^T \sqrt{n} \varphi(S_t, A_t) \varphi(S_t, A_t)^T \right]^T \rightarrow_p b^T W b 

II. \( n\rho_n \left[ \sum_t \zeta_{t+1}(\theta) \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \frac{1}{n\rho_n} \left[ \sum_t \zeta_{t+1}(\theta) \varphi(S_t, A_t)^T \right]^T \rightarrow_p \mathbb{E} \left[ \sum_t \psi_{t+1} \varphi(S_t, A_t)^T \right] W^{-1} \\
\times \mathbb{E} \left[ \sum_t \psi_{t+1} \varphi(S_t, A_t)^T \right]^T 

III. \( -2n\rho_n \left[ \sum_t \delta_{t+1} \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \frac{1}{n\rho_n} \left[ \sum_t b^T \sqrt{n} \varphi(S_t, A_t) \varphi(S_t, A_t)^T \right]^T \rightarrow_d -2Z_{\infty} b 

IV. \( 2n\rho_n \left[ \sum_t \delta_{t+1} \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \frac{1}{n\rho_n} \left[ \sum_t \zeta_{t+1}(\theta) \varphi(S_t, A_t)^T \right]^T \rightarrow_d 2Z_{\infty} W^{-1} \mathbb{E} \left[ \sum_t \psi_{t+1} \varphi(S_t, A_t)^T \right]^T 

V. \( -2n\rho_n \left[ \sum_t b^T \sqrt{n} \varphi(S_t, A_t) \varphi(S_t, A_t)^T \right] \tilde{W}^{-1} \frac{1}{n\rho_n} \left[ \sum_t \zeta_{t+1}(\theta) \varphi(S_t, A_t)^T \right]^T \rightarrow_p -2b^T \mathbb{E} \left[ \sum_t \psi_{t+1} \varphi(S_t, A_t)^T \right]^T 

where \( \zeta_{t+1}(\theta) = -\gamma \max_a \theta_0^T \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^T \varphi(S_{t+1}, a) \). The first part follows from a low of large numbers. Here, we prove part II and the rest follow similarly.

By adding and subtracting \( \gamma \max_{a \in \pi^*} b^T \sqrt{n} \varphi(S_{t+1}, a) \) to \( \zeta_{t+1}(\theta) \), we have

\[
\zeta_{t+1}(\theta) = -\gamma \max_{a \in \pi^*} (\theta_0 + b/\sqrt{n})^T \varphi(S_{t+1}, a) + \gamma \max_{a \in \pi^*} (\theta_0 + b/\sqrt{n})^T \varphi(S_{t+1}, a) + \gamma \max_{a \in \pi^*} b^T \sqrt{n} \varphi(S_{t+1}, a) 
\]

Thus, by Lemma 2, when \( \|b/\sqrt{n}\| \to 0 \) as \( n \to \infty \), we have

\[
\sqrt{n}[\zeta_{t+1}(\theta)] \to \psi_{t+1} 
\]

where \( \psi_{t+1} = \gamma \max_{a \in \pi^*} b^T \varphi(S_{t+1}, a) \) and by low of large numbers

\[
\sqrt{n}\rho_n \left[ \sum_t \zeta_{t+1}(\theta) \varphi(S_t, A_t)^T \right] \rightarrow_p \mathbb{E} \left[ \sum_t \psi_{t+1} \varphi(S_t, A_t)^T \right] 
\]

which completes the proof of part II.
We just showed that $\hat{V}(b)$ converges in distribution to

$$V(b) = -2Z_\infty b + 2Z_\infty W^{-1}\mathbb{E}\left[\sum_t \psi_{t+1} \varphi(S_t, A_t)^\top\right]^\top + b^\top W b - 2 b^\top \mathbb{E}\left[\sum_t \psi_{t+1} \varphi(S_t, A_t)^\top\right]^\top$$

$$\quad + \mathbb{E}\left[\sum_t \psi_{t+1} \varphi(S_t, A_t)^\top\right] W^{-1} \mathbb{E}\left[\sum_t \psi_{t+1} \varphi(S_t, A_t)^\top\right]^\top.$$

By assuming that $V(b)$ is uniquely minimized in $b$ and by continuity of $V(b)$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \arg\min_b \hat{V}(b) \to_d \arg\min_b V(b),$$

Note that when $\gamma = 0$, $\arg\min_b V(b) = Z_\infty W^{-1}$. Also, under assumption A.7, i.e., when $|\pi^*| = 1$, and $\gamma$ is small enough, $\arg\min_b V(b) = Z_\infty \Gamma$ where

$$\Gamma = \left[I - \gamma W^{-1}\mathbb{E}\left(\sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top\right)\right]^\top$$

$$\quad \times \left[W + \gamma^2 \mathbb{E}\left(\sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top\right) W^{-1} \mathbb{E}\left(\sum_t \varphi(S_{t+1}, \pi^*) \varphi(S_t, A_t)^\top\right)^\top\right]^{-1},$$

where $I$ is an identity matrix. Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \arg\min_b \hat{V}(b) \to_d N(0, \Gamma^\top \Sigma \Gamma),$$

with $\Sigma = \mathbb{E}\left[\left\{\sum_t \delta_{t+1} \varphi(S_t, A_t)^\top\right\}^\top \left\{\sum_t \delta_{t+1} \varphi(S_t, A_t)^\top\right\}\right]$. 

Appendix E. THE EFFECT OF THE CHOICE OF $\gamma$ ON THE ESTIMATED OPTIMAL REGIME.

In this section, we have estimated the optimal treatment regime under the simulation scenario discussed in the manuscript for different values of $\gamma$. For the smaller values of $\gamma$ ($\gamma = 0.1$),
Figure 5: Simulation: The effect of the choice of $\gamma$ on the estimated optimal regime.

the estimated optimal policy will be more myopic and does not suggest to augment any medication. This happens because the side effect out-weighs the treatment effect. However, as $\gamma$ gets larger the optimal policy suggests to augment more treatments simply because the long-term effect of treatments out-weighs the side effects. This shows that $\gamma$ balances the immediate and long-term effect of treatments. Results are presented in Figure 5

REFERENCES

Bather, J. (2000). Decision theory: an introduction to dynamic programming and sequential decisions, volume 180. Wiley Hoboken, NJ.
Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Collins, L., Murphy, S., and Bierman, K. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science*, 5(3):185–196.

Goldberg, Y. and Kosorok, M. R. (2011). Q-learning with censored data. *Annals of Statistics*, 40(1):529–560.

Grundy, S., Cleeman, J., Bairey Merz, C., Brewer Jr, H., Clark, L., Hunninghake, D., Pasternak, R., Smith Jr, S., Stone, N., et al. (2004). Implications of recent clinical trials for the national cholesterol education program adult treatment panel iii guidelines. *Journal of the American College of Cardiology*, 44(3):720–732.

Hunt, D. (2008). American diabetes association (ada) standards of medical care in diabetes 2008. *Diabetes Care*, 31:S12–S54.

Jordan, M. (2002). *An introduction to probabilistic graphical models*. University of California, Berkeley.

Kahn, S. E., Haffner, S. M., Heise, M. A., Herman, W. H., Holman, R. R., Jones, N. P., Kravitz, B. G., Lachin, J. M., O’Neill, M. C., Zinman, B., et al. (2006). Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *New England Journal of Medicine*, 355(23):2427–2443.

Laber, E., Qian, M., Lizotte, D. J., and Murphy, S. A. (2010). Statistical inference in dynamic treatment regimes. *arXiv preprint arXiv:1006.5831*. 

35
Lavori, P. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38.

Maei, H., Szepesvári, C., Bhatnagar, S., and Sutton, R. (2010). Toward off-policy learning control with function approximation. *Proc. ICML 2010*, pages 719–726.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322.

Moodie, E., Richardson, T., and Stephens, D. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural computation*, 1(2):281–294.

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S., Oslin, D., Rush, A., and Zhu, J. (2006). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*, 32(2):257–262.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., and Murphy, S. A. (2012). Q-learning: A data analysis method for constructing adaptive interventions. *Psychological methods*, 17(4):478.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. Translation of excerpts by D. Dabrowska and T. Speed. *Statistical Science*, 6:462–47.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. (2008). An analysis of linear models, linear value-function approximation, and feature selection for reinforcement
learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 752–759.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. (1987). 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 & Mathematics with Applications*, 14(9-12):923–945.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326. Springer: New York.

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine*, 27(23):4678–4721.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, 23(8):2379–2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, in press.
Si, J. (2004). *Handbook of learning and approximate dynamic programming*, volume 2. Wiley-IEEE Press.

Simester, D. I., Sun, P., and Tsitsiklis, J. N. (2006). Dynamic catalog mailing policies. *Management Science*, 52(5):683–696.

Sutton, R. and Barto, A. (1998). *Reinforcement learning: An introduction*, volume 28. Cambridge Univ Press.

Sutton, R., Maei, H., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 993–1000. ACM.

Sutton, R., Szepesvári, C., and Maei, H. (2009b). A convergent $o(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in Neural Information Processing Systems 21*.

Tang, Y. and Kosorok, M. R. (2012). Developing adaptive personalized therapy for cystic fibrosis using reinforcement learning. Technical report, The University of North Carolina at Chapel Hill.

Timbie, J., Hayward, R., and Vijan, S. (2010). Diminishing efficacy of combination therapy, response-heterogeneity, and treatment intolerance limit the attainability of tight risk factor control in patients with diabetes. *Health Services Research*, 45(2):437–456.

Vapnik, V., Golowich, S. E., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pages 281–287.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.
Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(2):1–14.

ZHAO, Y., KOSOROK, M. R., and ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.