

Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings

BY ASHKAN ERTEFAIE

Department of Statistics, The Wharton school, University of Pennsylvania

3730 Walnut Street, Philadelphia, PA, USA

ertefaie@wharton.upenn.edu

SUMMARY

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 to mimic 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. Supplementary materials for this article are available online.

Some key words: 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 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 strategy (Lavori & Dawson, 2000).

The goal of this manuscript is to devise a new methodology that can be used to construct the optimal DTR in infinite-horizon settings (i.e., when the number of decision points is not necessarily fixed for all individuals). The estimation procedure, however, is based on observational data collected over a fixed period of time that 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), which was designed to assess the health status of adults and children in 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 ability of available treatments to control risk factors. Specifically, Timbie and colleagues' study was designed to manage the risk factors for vascular complications such as high blood pressure, high cholesterol and high 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 challenge in constructing an optimal regime is avoiding 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 and Robins' methods 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 & 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). 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 with data that 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 capture the disease dynamic and the long-term treatment effects, our dataset 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 the method. We conduct a simulation study in Section 4 to examine the performance of our method. The last section contains some concluding remarks. All the proofs are relegated to an online supplementary document.

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 $(X_{i0}, A_{i0}, \dots, A_{i(T-1)}, X_{iT})$, 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 the observed

history (such as time-varying covariates, prior response, baseline covariates and treatment history) that depends on, at most, the last k time points. We assume that the support of S_t is the same for all t s and denote it as \mathcal{S} . If a patient dies before the last decision point, say t , we set $S_t = \emptyset$ (absorbing state). Given $S_t = s$, A_t takes values in $\mathcal{A}_s = \{0, 1, 2, \dots, m_s\}$ for all t where $m_s < \infty$, $\forall s \in \mathcal{S}$. We set $A_s = \emptyset$ for $s = \emptyset$. The treatment and the summary function history through t are denoted by \bar{A}_t and \bar{S}_t , respectively. We use lowercase 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 outcomes models were 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 $\bar{S}_{m+1}(\bar{a}_m) = (S_2(a_1), \dots, S_{m+1}(\bar{a}_m))$, 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 outcomes 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*: $\{S_{t+1}(\bar{a}_t), S_{t+2}(\bar{a}_{t+1}), \dots, S_T(\bar{a}_{T-1})\} \perp\!\!\!\perp A_t | \bar{S}_t, \bar{A}_{t-1} = \bar{a}_{t-1}$.

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 ,

$$S_t \perp\!\!\!\perp \bar{S}_{t-2}, \bar{A}_{t-2} | S_{t-1}, A_{t-1} \quad (1)$$

$$A_t \perp\!\!\!\perp \bar{S}_{t-1}, \bar{A}_{t-1} | S_t \quad (2)$$

A.2 Time homogeneity: For each $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$,

$$p(S_{t+1} \in \mathcal{B} | S_t = s, A_t = a) = p(S' \in \mathcal{B} | S = s, A = a),$$

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 $a \in \mathcal{A}_s$ and for each possible value s , $p_{A|S}(a|s) > 0$.

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 \mathcal{A}_s have been assigned to some patients (i.e., for each $S = s$, all actions in \mathcal{A}_s are possible). This assumption is also known as an *exploration* assumption.

Assumptions A.1 and A.2 provide guidance for how to construct the state variable S_t . The Markovian assumptions (1) and (2) seem to be unrealistic in many studies of chronic diseases.

But they are not. This is because, if it is necessary, one can construct the state variable S_t such that it includes previous treatments and observed intermediate outcomes. Thus, for example, (2) does *not* indicate that decision makers make the next treatment decision taking into account only the last outcome data, i.e. disregarding the earlier treatments and outcomes, because these information can be included in the preceding state variable, say S_t .

In cases where S_t has to depend on the observed history of the last k time points, assumption A.2 is satisfied only if we ignore the first $k - 1$ time points of the observed trajectory of patients. This is because the support of S_t is the same only for $t \geq k$.

2.3. The likelihood

Under assumptions A.1–3, the distribution of the observed trajectories is composed of the distribution of the trajectory S_t given (S_{t-1}, A_{t-1}) , say $f_{S'|S,A}$, and the density $p_{A|S}(a|s)$. The likelihood of the observed trajectory $\{s_0, a_0, \dots, a_{T-1}, s_T\}$ is given by

$$f_S(s_0)p(a_0|s_0) \prod_{k=1}^T f_{S'|S,A}(s_k|s_{k-1}, a_{k-1}) \prod_{k=1}^{T-1} p(a_k|s_k). \quad (3)$$

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

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

$$f(s_0)I(a_0 = \pi(s_0)) \prod_{k=1}^T f_{S'|S,A}(s_k|s_{k-1}, a_{k-1}) \prod_{k=1}^{T-1} I(a_k = \pi(s_k)). \quad (4)$$

Expectations with respect to this distribution are denoted by \mathbb{E}_π . Note the likelihood (4) is not well-defined and it may be identical to zero *unless* A.3 holds for each possible value s and $a \in \mathcal{A}_s$. Note that, the observed trajectory $\{s_0, a_0, \dots, a_{T-1}, s_T\}$ may be truncated by death at time m . In this case, we have $S_m = S_{m+1} = \dots = S_T = \emptyset$, and $A_m = A_{m+1} = \dots = A_T = \emptyset$ and by definition, for all $m' \geq m$, $p(S_{m'+1} = \emptyset | S_{m'} = \emptyset, A_{m'} = \emptyset) = 1$ and $p(A_{m'} = \emptyset | S_{m'} = \emptyset) = 1$.

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 , $Q_t^\pi(s, a)$, is defined as an expected value of the cumulative discounted reward if taking treatment a at state s at time t and following the policy π afterward. In other words, $Q_t^\pi(s, a)$ quantifies the quality of policy π when $S_t = s$ and $A_t = a$. Hence, $Q_t^\pi(s, a)$ is defined as $\mathbb{E}_\pi [\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} | S_t = s, A_t = a]$, where γ is called a discount factor, $0 < \gamma < 1$, which is fixed a priori by the researcher. Note that by definition of the reward function, $Q_t^\pi(\emptyset, a) = 0$. 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 γ *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 γ approaches one, future rewards become more important. In other words, γ specifies our inferential goal. In Section S4 of the supplementary materials we discuss the effect of the choice of γ on the estimated optimal regime.

The action-value function can be written as

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

The last equation is known as *Bellman equation* for $Q^\pi(s, a)$ (Sutton & Barto, 1998; Si, 2004). The inner expectation quantifies the quality of policy π 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 π is followed in the interval $(t, t+1]$.

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

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

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

$$Q^*(s, a) = \mathbb{E} \left[R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') | S_t = s, A_t = a \right]. \quad (5)$$

Note that the only distribution involved in the \mathbb{E} is $f_{S'|S, A}$. Denote any policy π^* 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. So, for state s , we can define the optimal policy as $\pi^*(s) = \arg \max_a Q^*(s, a)$ and the optimal value function as $V^*(s) = Q^*(s, \pi^*(s))$.

The action-value function $Q^*(s, a)$ can be estimated by turning the recurrence relation (5) into an update rule that relies on estimating the conditional density $f_{S'|S, A}$. However, when the cardinality of $(\mathcal{S}, \mathcal{A})$ and the dimension of S_t are large, estimation of the conditional densities is 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 a 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 (5) 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 θ_0 ,

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

where θ_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 θ_0 . First, we define an error term and then we construct an estimating equation. In view of the Bellman equation (5), for each t , we have

$$\mathbb{E} \left[R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') - Q^*(S_t, A_t) | S_t = s, A_t = a \right] = 0. \quad (6)$$

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 θ s, we multiply $\delta(\theta)$ by $\varphi(S, A)$ and define θ_0 as a value of θ such that

$$D(\theta) = \mathbb{E} \left[\sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right] = 0. \quad (7)$$

The expectation in the above equation depends on the transition densities $f_{S'|S,A}$ and $p_{A|S}$. Note that as in (6),

$$\begin{aligned} D(\theta_0) &= \mathbb{E} \left[\sum_{t=0}^{T-1} \{R_{t+1} + \gamma \max_a [\theta_0^\top \varphi(S_{t+1}, a)] - \theta_0^\top \varphi(S_t, A_t)\} \varphi(S_t, A_t)^\top \right] \\ &= \sum_{t=0}^{T-1} \mathbb{E} \left[\{R_{t+1} + \gamma \max_a [\theta_0^\top \varphi(S_{t+1}, a)] - \theta_0^\top \varphi(S_t, A_t)\} \varphi(S_t, A_t)^\top \right] \\ &= 0. \end{aligned}$$

Hence, given the observed data, an unbiased estimating equation for θ can be defined as

$$\hat{D}(\theta) = \mathbb{P}_n \left[\sum_{t=0}^{T-1} \delta_{t+1}(\theta) \varphi(S_t, A_t)^\top \right] = 0, \quad (8)$$

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

3. CALCULATION

In practice, sometimes there is no $\hat{\theta}$ that solves the system of equations (8), 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 (7), a simple objective

function can be defined as

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

(Sutton et al., 2009b). The objective function used in this manuscript is the above function weighted by the inverse of the feature covariance matrix and defined as

$$M(\theta) = D(\theta) W^{-1} D(\theta)^\top, \quad (9)$$

where $W = \mathbb{E} \left[\sum_{t=0}^{T-1} \varphi(S_t, A_t) \varphi(S_t, A_t)^\top \right]$ is a full-rank matrix. The weight W^{-1} improves the performance of the proposed stochastic minimization algorithm in Section S1 of the supplementary materials. 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

$$\hat{M}(\theta) = \hat{D}(\theta) \hat{W}^{-1} \hat{D}(\theta)^\top, \quad (10)$$

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 $\hat{\pi}(s) = \arg \max_a \hat{\theta}^\top \varphi(s, a)$. By law of large numbers, the estimator $\hat{M}(\theta)$ is a consistent estimator of $M(\theta)$.

The following theorem presents the consistency and asymptotic normality of estimator $\hat{\theta}$ where the asymptotic normality result relies on the uniqueness of the optimal treatment at each decision point. This allows investigators to test the significance of variables for use in this sequential decision making problem. Assumptions A.4 – 8 required in this section are listed in Appendix 3.

THEOREM 1. (*Consistency and asymptotic normality*) For a map $\theta \rightarrow M(\theta)$, defined in (9), under assumptions A.4 – 7, any sequence of estimators $\hat{\theta}$ with $\hat{M}(\hat{\theta}) \leq \hat{M}(\theta_0) + o_p(1)$ satisfies the following statements:

- I. For small enough γ , $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$.
- II. Under A.8, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Gamma^\top \Sigma \Gamma)$, where

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

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

Proof. See the supplementary material. \square

In Section S3 of the supplementary materials, we show that under assumption A.8, $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ for any $\gamma \in (0, 1)$. The asymptotic variance $\Gamma^\top \Sigma \Gamma$ may be estimated consistently by replacing the expectations with expectations with respect to the empirical measure and replacing θ_0 with its estimate $\hat{\theta}$.

The objective function $M(\theta)$ is a non-convex and non-differentiable function of θ , which complicates the estimation process. Standard optimization techniques often fail to find the global minimizer of this function. In Section S1 of the supplementary materials, we present an incremental approach, which is a generalization of greedy gradient Q-learning (GGQ), an iterative stochastic minimization algorithm, introduced by Maei et al. (2010) as a tool to minimize $M(\theta)$. Hence, from this point forward, we refer to our proposed method as GGQ.

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. These simulated data mimic the third wave of NHANES.

In Section 4.3, we compare the performance of our proposed approach with the *classical* approach using a simulated dataset and then in Section 4.4 we perform a Monte Carlo study to examine the asymptotic results.

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 treatment at time t . As soon as a treatment is augmented the variable NAT increases by one, whether or not the treatment will be discontinued. The death indicator variable C_t at time t is generated as a function of previous observed covariates.

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 a multivariate normal distribution with mean $(13, 160, 9.4)$ and the covariance matrix $diag(1, 1, 1)$, where BP is the systolic blood pressure. Also, $NAT_0 = D_0 = C_0 = 0$.
- *Assigned treatment at time t :* Given the state variable NAT_t , the sets of available treatments are $\mathcal{A}_{NAT_t=0} = \{0, Metformin\}$, $\mathcal{A}_{NAT_t=1} = \{0, Sulfonylurea\}$, $\mathcal{A}_{NAT_t=2} = \{0, Glitazone\}$, $\mathcal{A}_{NAT_t=3} = \{0, Insulin\}$, and $\mathcal{A}_{NAT_t=4} = \{0\}$, 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 $A_t = insulin$ 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, $\mathbb{E}[A1c_t|Z_t = 1, A1c_{t-1}] = \mathbb{E}[A1c_t|A_t = 0, A1c_{t-1}] = A1c_{t-1}$.

- *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 $p(D_t|A_{t-1} = metformin) = p(D_t|A_{t-1} = sulfonylurea) = p(D_t|A_{t-1} = glitazone) = 0.20$, and $p(D_t|A_{t-1} = insulin) = 0.35$.

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

- *Intermediate outcome A1c at time t:* To avoid variance inflation through time, we use the following generative model for $A1c$ at time t , $A1c_t = \frac{A1c_{t-1}-\mu_{t-1}+\epsilon}{\sqrt{(1+\sigma_\epsilon^2)}} + \mu_t$, where $\epsilon \sim N(0, \sigma_\epsilon = 0.5)$ and

$$\mu_t = \mathbb{E}[A1c_t|A1c_{t-1}, NAT_{t-1}, A_t, D_t] = \begin{cases} \mu_{t-1}(1 - \tau_{A_t}) & \text{if } A1c_{t-1} > 7, NAT_{t-1} < 4, A_t \neq 0, D_t \neq 1, \\ \mu_{t-1} & \text{o.w.} \end{cases}$$

with τ_{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 $A1c$. 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:* $BP_t = (BP_{t-1} + \epsilon)/(\sqrt{1 + \sigma_\epsilon^2})$ and $Weight_t = (Weight_{t-1} + \epsilon)/(\sqrt{1 + \sigma_\epsilon^2})$.
- *Death indicator at time t:* A binary variable C_t is generated from a Bernoulli distribution with probability $\frac{\exp\{-10+0.08I(A1c_{t-1}>7)A1c_{t-1}^2+0.5NAT_{t-1}\}}{1+\exp\{-10+0.08I(A1c_{t-1}>7)A1c_{t-1}^2+0.5NAT_{t-1}\}}$. $C_t = 1$ is the indicator of death.
- *Reward function at time t:* In order to be able to find an optimal treatment regime, we need an operational definition of controlled $A1c$. Hence, we define the following reward function at time t as a function of $A1c_t$, D_t and C_t ,
 - $R_t = 1$ if $A1c_t < 7$, -2 if $7 < A1c_t \& D_t = 1$, -10 if $C_t = 1$ and zero otherwise.

This reward structure helps us to identify treatments whose discontinuation rate outweighs their efficacy while reducing the chance of death.

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. In order to satisfy assumption A.2, we ignored the first four time points in the observed trajectory of each patient.

4.3. Analysis of a simulated dataset

We generate two datasets of sizes 2,000 and 5,000 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 (5) into an update rule as

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

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

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

The above 5-step procedure is similar to the one presented in Chapter 4 of Sutton & 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 small. We categorize the continuous variables ($BP, Weight, A1c$) and estimate $P_{S'|S,A}$ nonparametrically. 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 $(-\infty, 7, 7.2, 7.5, 7.7, 8, 9, +\infty)$. 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 $S_t = (NAT_t, D_t, A1c_t, BP_t, Weight_t)$. In our example, we parametrize the optimal action value function $Q^*(s, a)$ using a 72-dimensional vector of parameters and construct the features $\varphi(s, a)$ using radial basis functions (Gaussian kernels). See Appendix 2 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 1 and multiply them by $v \in (0, 1)$. Then we run the algorithm for different values of v and select the one that minimizes the objective function. In this simulation study we set the step sizes $\alpha_k = \nu / (k \log(k))$ and $\beta_k = \nu / k$ where ν is set to 0.05. Section S5 in the supplementary materials discusses the effect of the choice of tuning parameters on the estimated optimal regime.

True optimal policy. As the sample size 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 ($NAT_t, D_t, Cat.A1c_t$). Then by the 5-step procedure (classical approach), the true optimal policy is approximated and set as our benchmark.

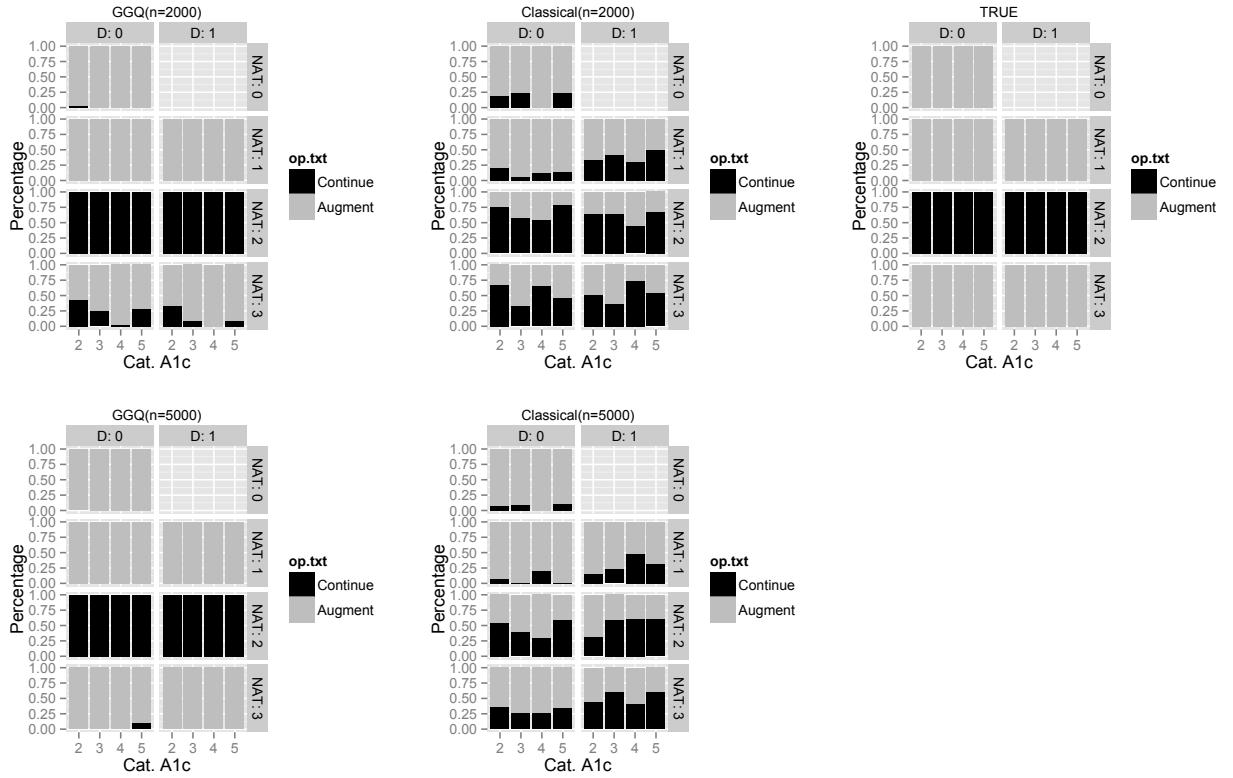


Fig. 1: Simulation: Estimated optimal treatment (op.txt) for different states. The shaded bar represents the evidence in the simulated data for each of the treatment choices as labeled in the legend. The upper and lower horizontal axes are the discontinuation indicator and the categorized A1c (Cat.A1c), 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.

Figure 1 depicts the true and estimated optimal treatment for each discretized oracle state ($NAT_t, D_t, Cat.A1c_t$) using the GQ and classical approaches. As in this example, we set the discount factor γ to 0.6. Note that the estimated optimal policy using classical and GQ methods is based on the states S_t^{Class} and S_t , 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 axis on the left hand side is the percentage of time that the treatment choices are selected as optimal. The left vertical and both horizontal axes represent the elements of the state ($NAT, D, Cat.A1c$), respectively. This plot shows that the proposed GQ method outperforms the classical approach for moderate sample sizes.

More specifically, the estimated optimal policy using GQ ($\hat{\pi}_{GQ}$) 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 outweighs 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 $\hat{\pi}_{GQ}$ and $\hat{\pi}_{Class}$. The values are calculated using the Monte Carlo method, where the value of a treatment

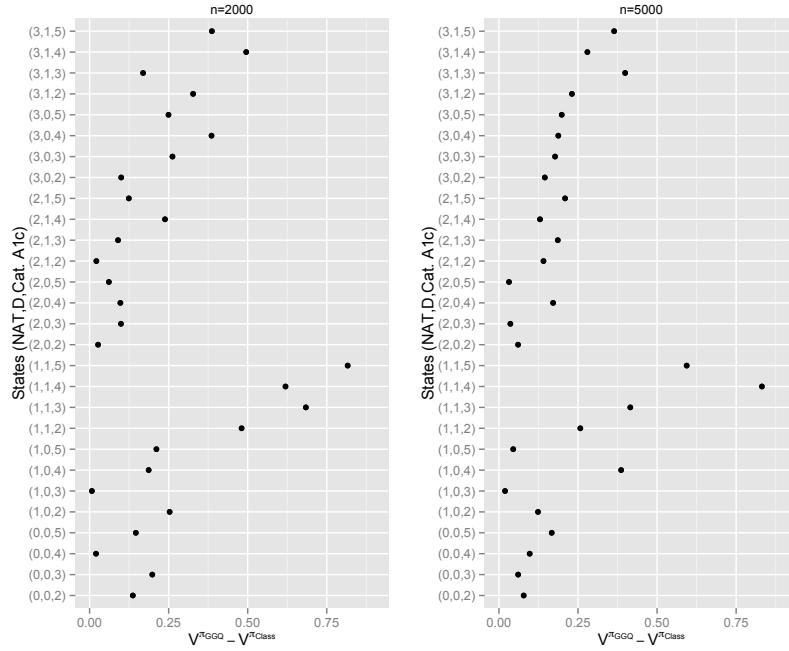


Fig. 2: Simulation: Monte Carlo approximation of the difference between value functions. $V^{\pi_{GGQ}}$ and $V^{\pi_{Class}}$ are the value functions corresponding to the classical and GGQ approaches. The vertical axis represents the triplets of states in the order of $(NAT, D, Cat. A1c)$.

policy π , 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.

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.

4.4. Monte Carlo studies

We generate 500 datasets each of sizes 2,000 and 5,000 to examine the asymptotic behavior of the proposed method. Figure 3 shows the confidence intervals of θ , where the standard errors are estimated using the variance formula presented in Theorem 1. The dark circles are the average of the 500 $\hat{\theta}$'s and asterisks are the true parameter values approximated using a Monte Carlo study with $n = 10,000$. These confidence intervals may be used to identify the parts of the feature function that should be kept in the decision rule and may be used as a feature selection tool. Another important use of the asymptotic results in Theorem 1 is to investigate whether there is a significant difference between treatment options. Specifically, one may build the confidence interval for the difference between the estimated optimal action-value function for different treatment options (augment vs. continue) and check whether it contains zero (while adjusting for Type-I error rate for more than two treatment options). In Figure 4, we constructed and evaluated the quality of these 95% confidence intervals for $Q^*(s, \text{augment}) - Q^*(s, 0)$, given each state variable for $n = 2000$. The results for $n = 5000$ is similar and omitted due to space limitations. These results confirm the accuracy of our variance estimator in Theorem 1.

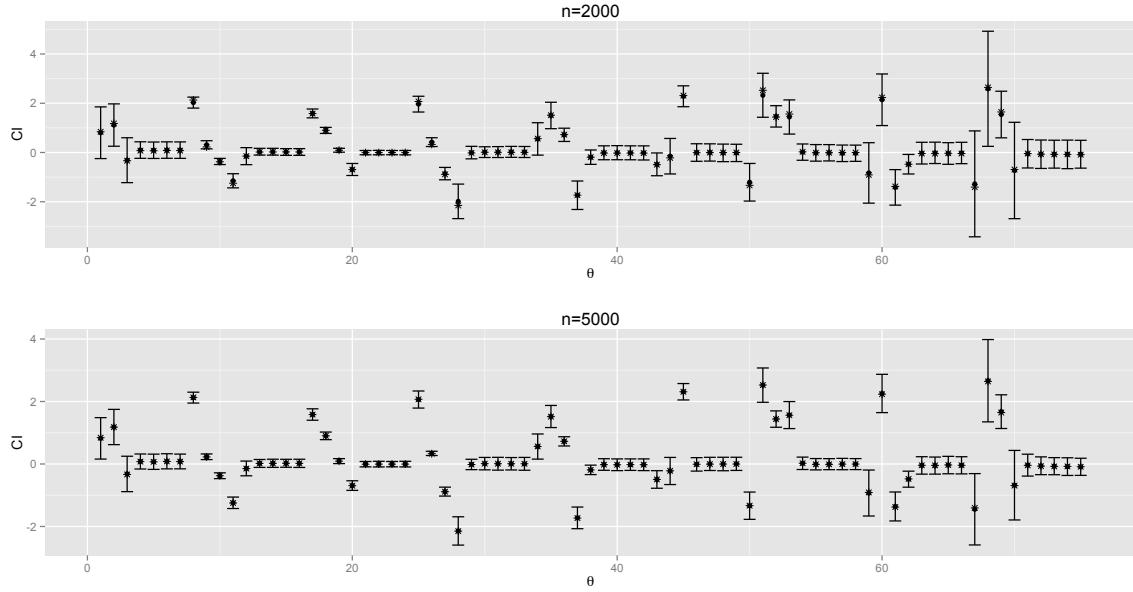


Fig. 3: Simulation: Plot of the 95% confidence intervals of θ .

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 estimators 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.8 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 & Darken, 1989; Poggio & Girosi, 1990). One simple method is to try different feature functions (φ) and select the one that minimizes the function $f(\varphi) = \min_{\theta} M(\theta)$ (Parr et al., 2008). Alternatively, one may use support vector regression to approximate the action-value function (Vapnik et al., 1997; Tang & 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). 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

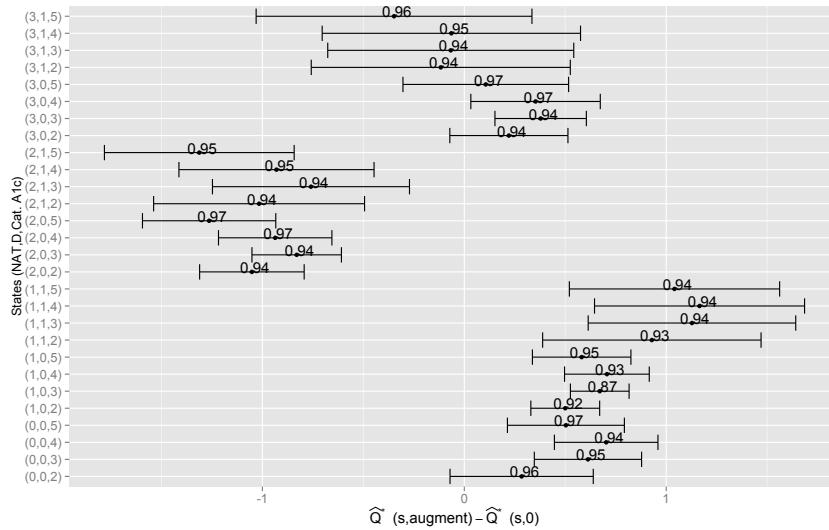


Fig. 4: Simulation: The confidence intervals of the difference between the estimated optimal action-value function when the treatment is augmented and continued (i.e., $\hat{Q}^*(s, \text{augment}) - \hat{Q}^*(s, 0)$). The number on each confidence interval represents the coverage of that interval. The vertical axis represents the triplets of states in the order of (*NAT, D, Cat.A1c*).

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

Acknowledgements should appear after the body of the paper but before any appendices and be as brief as possible subject to politeness. Information, such as contract numbers, of no interest to readers, must be excluded.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the stochastic minimization algorithm and proof of Theorem 1. It also discusses the effect of the discount factor and tuning parameters of the stochastic minimization algorithm on the estimated optimal treatment regime.

APPENDIX 1: TUNING PARAMETERS

The tuning parameters α_k and β_k in the GGQ algorithm need to satisfy the following assumptions (Maei et al., 2010):

- 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 2: FEATURE FUNCTIONS

The feature functions are constructed using the radial basis functions and

$$\varphi(s, a) = I(s \neq \emptyset)(\varphi_1(s, a), \varphi_2(s, a), \varphi_3(s, a), \varphi_4(s, a), \varphi_5(s, a), \varphi_6(s, a), \varphi_7(s, a), \varphi_8(s, a), \varphi_9(s, a)),$$

where

$$\begin{aligned} \varphi_1(s, a) &= I(A = 0, NAT = 0)(1, \exp[-h(A1c - q_{11})^2], \exp[-h(A1c - q_{12})^2], \phi(BP), \phi(Weight)) \\ \varphi_2(s, a) &= I(A = 0, NAT = 1)(1, \exp[-h(A1c - q_{21})^2], \exp[-h(A1c - q_{22})^2], \exp[-h(A1c - 8.0)^2], \\ &\quad d, \phi(BP), \phi(Weight)) \\ \varphi_3(s, a) &= I(A = 0, NAT = 2)(1, \exp[-h(A1c - q_{31})^2], \exp[-h(A1c - q_{33})^2], d, \phi(BP), \phi(Weight)) \\ \varphi_4(s, a) &= I(A = 0, NAT = 3)(1, \exp[-h(A1c - q_{41})^2], \exp[-h(A1c - q_{42})^2], \exp[-h(A1c - 8.5)^2], \\ &\quad d, \phi(BP), \phi(Weight)) \\ \varphi_5(s, a) &= I(A = 0, NAT = 4)(1, \exp[-h(A1c - q_{51})^2], \exp[-h(A1c - q_{52})^2], \exp[-h(A1c - 8.0)^2], \\ &\quad d, \phi(BP), \phi(Weight)) \\ \varphi_6(s, a) &= I(A = 1, NAT = 0)(1, \exp[-h(A1c - 6.5)^2], \exp[-h(A1c - 7.5)^2], \phi(BP), \phi(Weight)) \\ \varphi_7(s, a) &= I(A = 2, NAT = 1)(1, \exp[-h(A1c - 6.5)^2], \exp[-h(A1c - q_{71})^2], \exp[-h(A1c - q_{73})^2] \\ &\quad d, \phi(BP), \phi(Weight)) \\ \varphi_8(s, a) &= I(A = 3, NAT = 2)(1, \exp[-h(A1c - q_{82})^2], \exp[-h(A1c - 8.5)^2], d, \phi(BP), \phi(Weight)) \\ \varphi_9(s, a) &= I(A = 4, NAT = 3)(1, \exp[-h(A1c - 6.5)^2], \exp[-h(A1c - q_{92})^2], \exp[-h(A1c - 8.5)^2], \\ &\quad d, \phi(BP), \phi(Weight)), \end{aligned}$$

where $\phi(BP) = (\exp[-h(BP - q_{b1})^2], \exp[-h(BP - q_{b3})^2])$ 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_{b1} and q_{b3} 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 quantiles used in each φ_k and $\phi(\cdot)$ is a bias-variance trade-off such that increasing the number of quantiles 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 3: ASSUMPTIONS

In addition to assumptions A.1-3, the following assumptions are required for large sample properties of our estimator.

A.4 $\theta_0^\top \varphi(\cdot, \cdot)$ is the optimal Q-function.

A.5 $\mathbb{E} \left[\sum_{t=0}^{T-1} \|\varphi(S_t, A_t)\|_2 \|\varphi(S_t, a)\|_2 \right] < \infty$, for any $a \in \mathcal{A}$.

A.6 The matrix W is of full rank.

A.7 $\mathbb{E} \left[\sum_{t=0}^{T-1} \{\gamma I_{|\pi^*(S_{t+1})|=1} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t)\} \varphi(S_t, A_t)^\top \right]$ is of full rank where $\pi^*(S_{t+1}) = \arg \max_a \theta_0^\top \varphi(S_{t+1}, a)$ and $|\cdot|$ is the cardinal of a set.

A.8 The optimal treatment is unique at each decision point.

**Web-based Supplementary Materials for
“Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings”**

S1. THE STOCHASTIC MINIMIZATION ALGORITHM

We base our minimization procedure on the (approximate) gradient descent approach, in which subgradients are defined as Frechet sub-gradients of the objective function $M(\theta)$. Following Maei et al. (2010) and under assumptions A.1-8, the algorithm converges to the minimizer of our objective function.

The sub-gradient $\partial M(\theta)$ of $M(\theta)$ with respect to θ is

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

where $\varpi = \mathbb{E}[\sum_t \varphi(s_t, a_t) \varphi(s_t, a_t)^\top]^{-1} \mathbb{E}[\sum_t \delta_{t+1}(\theta) \varphi(s_t, a_t)^\top]$. 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 θ and ϖ . Using grid search, the initial value θ_1 can be set as the one that minimizes the objective function and the initial value $w_1 = \mathbb{P}_n[\sum_t \varphi(s_t, a_t) \varphi(s_t, a_t)^\top]^{-1} \mathbb{P}_n[\sum_t \delta_{t+1}(\theta_0) \varphi(s_t, a_t)^\top]$.
2. Start from the first individual's trajectory and obtain θ_{k+1} from the following iterative equations:

$$\theta_{k+1} = \theta_k + \alpha_k \nu \sum_t [\delta_{t+1}(\theta_k) \varphi(s_t, a_t) - \gamma \{\varpi_k^\top \varphi(s_t, a_t)\} \varphi(s_{t+1}, \pi_{\theta_k}^*(s_{t+1}))^\top] \quad (\text{C1})$$

$$\varpi_{k+1} = \varpi_k + \beta_k \nu \sum_t [\delta_{t+1}(\theta_k) - \{\varphi(s_t, a_t)^\top \varpi_k\}^\top] \varphi(s_t, a_t)^\top, \quad (\text{C2})$$

where α_k , β_k and ν are tuning parameters (step sizes) and $\pi_{\theta_k}^*(.)$ is the optimal policy estimated as a function of θ_k .

3. Use step 2 to continue updating the parameters to the last individual.
4. Continue steps 2 and 3 until $\|\theta_{k+1} - \theta_k\|_2 < c$ where c is a constant.

The tuning parameters (step sizes) α_k and β_k need to satisfy assumptions P.1-P.4 in Appendix A. The parameter ν tunes the step sizes and lies in the interval (0,1). Our simulation studies show that the best choice of ν would be close to $\frac{1}{T}$, where T is the maximal length of the trajectories in our data. See our discussion in Section S5.

S2. LEMMA

LEMMA 1. Let $\{a_i\}_1^K$ and $\{b_i\}_1^K$ be two sets of elements, then

- I. $\lim_{\|b\| \rightarrow 0} \frac{\max_{1 \leq i \leq K} [a_i + b_i] - \max_{i \in \pi^*} [a_i + b_i]}{\|b\|} = 0$;
- 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

$$\max_{i \notin \pi^*} b_i - \max_{i \in \pi^*} b_i \leq \max_{1 \leq i \leq K} b_i - \max_{i \in \pi^*} b_i$$

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

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

$$\begin{aligned} 0 &\leq \frac{\max_{1 \leq i \leq K} [a_i + b_i] - \max_{i \in \pi^*} [a_i + b_i]}{\|b\|} \\ &= \max_{1 \leq i \leq K} \left[a_i - a_{i^*} + b_i - \max_{i \in \pi^*} b_i \right] \frac{1}{\|b\|}, \quad \forall i^* \in \pi^* \\ &= \max \left[\max_{i \in \pi^*} \{b_i - \max_{i \in \pi^*} b_i\}, \max_{i \notin \pi^*} \{a_i - a_{i^*} + b_i - \max_{i \in \pi^*} b_i\} \right] \frac{1}{\|b\|} \\ &= \max \left[0, \max_{i \notin \pi^*} \left\{ \frac{a_i - a_{i^*}}{\|b\|} + \frac{b_i - \max_{i \in \pi^*} b_i}{\|b\|} \right\} \right] \\ &\leq \max \left[0, \max_{i \notin \pi^*} \left\{ \frac{a_i - a_{i^*}}{\|b\|} + \frac{2 \max_{1 \leq i \leq K} b_i}{\|b\|} \right\} \right] \end{aligned}$$

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.

S3. PROOF OF THEOREM 1

We first show that the objective function $M(\theta)$ is continuous. Then using the results of Lemma 1, we show that $M(\theta)$ satisfies the two required conditions of Theorem 3.2.1 in Van Der Vaart & Wellner (1996), which completes the proof of consistency. To prove the asymptotic normality, under the additional assumption $A.8$, 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

$$\begin{aligned} \|D(\theta) - D(\theta_0)\| &= \left\| \sum_{t=0}^{T-1} \mathbb{E} \left[\left\{ \max_a [\theta^\top \varphi(S_{t+1}, a)] - \max_a [\theta_0^\top \varphi(S_{t+1}, a)] \right. \right. \right. \\ &\quad \left. \left. \left. - \theta^\top \varphi(S_t, A_t) + \theta_0^\top \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right] \right\|, \end{aligned}$$

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

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

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

Part I (Consistency). We show that $M(\theta)$ 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 $\|\theta - \theta_0\| < \epsilon$,

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

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

$$D(\theta)W^{-1}D(\theta)^\top \geq c\|\theta - \theta_0\|^2. \tag{C3}$$

The left hand side of the above inequality can be written as

$$\begin{aligned} D(\theta)W^{-1}D(\theta)^\top &= [D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]W^{-1}[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]^\top \\ &\quad + \dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}\dot{D}_{\theta_0}(\theta - \theta_0)^\top + 2\dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]^\top, \end{aligned}$$

where

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

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

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

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

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

Thus,

$$\begin{aligned} \frac{\|D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\|}{\|\theta - \theta_0\|} &= \\ \frac{\gamma}{\|\theta - \theta_0\|} \left\| \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ \max_a (\theta_0 + b\|\theta - \theta_0\|)^\top \varphi(S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} (\theta_0 + b\|\theta - \theta_0\|)^\top \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right] \right\| & \\ \leq \gamma \mathbb{E} \left[\sum_{t=0}^{T-1} \|\varphi(S_{t+1}, a)\| \left\{ \max_a \left(\frac{\theta_0}{\|\theta - \theta_0\|} + b \right)^\top \varphi(S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} \left(\frac{\theta_0}{\|\theta - \theta_0\|} + b \right)^\top \varphi(S_{t+1}, a) \right\} \right] & \\ \leq \gamma \mathbb{E} \left[\sum_{t=0}^{T-1} \|\varphi(S_t, A_t)\| \left\{ \max_a b^\top \varphi(S_{t+1}, a) - \max_{a \in \pi^*(S_{t+1})} b^\top \varphi(S_{t+1}, a) \right\} \right] & \\ \leq \gamma \mathbb{E} \left[\sum_{t=0}^{T-1} \|\varphi(S_t, A_t)\| \sum_{a \in \mathcal{A}_{S_{t+1}}} \|b\| \|\varphi(S_{t+1}, a)\| \right] < \infty. & \end{aligned}$$

The second inequality follows from Lemma 1 part (II) and the last inequality follows from $\|b\| = 1$ and assumption A.5. Also, using Lemma 1 part (I), we have

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

We just showed that $\|D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|)$. Since W^{-1} is of full rank matrix,

$$[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]W^{-1}[D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)]^\top = o(\|\theta - \theta_0\|^2).$$

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,

$$\begin{aligned}\dot{D}_{\theta_0}(\theta - \theta_0) &= \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ \gamma \max_{a \in \pi^*(S_{t+1})} (\theta - \theta_0)^\top \varphi(S_{t+1}, a) - (\theta - \theta_0)^\top \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right] \\ &= (\theta - \theta_0)^\top \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|=1} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right] \\ &\quad + \gamma \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|>1} \max_{a \in \pi^*(S_{t+1})} \frac{(\theta - \theta_0)^\top}{\|\theta - \theta_0\|} \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right] \|\theta - \theta_0\|.\end{aligned}$$

Let

$$\begin{aligned}M_1 &= \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ \gamma I_{|\pi^*(S_{t+1})|=1} \varphi(S_{t+1}, \pi^*(S_{t+1})) - \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right], \\ M_2 &= \mathbb{E} \left[\sum_{t=0}^{T-1} \left\{ I_{|\pi^*(S_{t+1})|>1} \max_{a \in \pi^*(S_{t+1})} \frac{(\theta - \theta_0)^\top}{\|\theta - \theta_0\|} \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right].\end{aligned}$$

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.7), we have

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

where λ_{min} is the smallest eigenvalue of $M_1^\top M_1$. Also, using singular value decomposition we have

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

where λ_{max} is a maximum eigenvalue of $M_1 M_1^\top$. Thus

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

Therefore, function $M(\cdot)$ satisfies the first condition of Theorem 3.2.5 in Van Der Vaart and Wellner (1996) for any small enough γ such that $[\lambda_{min} - 2\gamma \sqrt{\lambda_{max}} \|M_2\| - \gamma^2 \|M_2\|^2] > 0$. Note that under assumption A.8 when $|\pi^*(S_{t+1})| = 1$, the latter condition is satisfied automatically. Since W is of full rank (Assumption A.6), $\dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}\dot{D}_{\theta_0}(\theta - \theta_0)^\top \geq c\|\theta - \theta_0\|^2$ for $c > 0$.

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

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

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

$$\begin{aligned}|[\hat{M}(\theta) - M(\theta)] - [\hat{M}(\theta_0) - M(\theta_0)]| &= |(\hat{D}(\theta) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta) - \hat{D}(\hat{\theta}))^\top \\ &\quad - (\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))\hat{W}^{-1}(\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))^\top \\ &\quad - (D(\theta) - D(\theta_0))\hat{W}^{-1}(D(\theta) - D(\theta_0))^\top|.\end{aligned}$$

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

$$\begin{aligned} \mathbb{E} \sup_{\|\theta - \hat{\theta}\| \leq \delta_n} [(\hat{D}(\theta) - \hat{D}(\hat{\theta})) \hat{W}^{-1} (\hat{D}(\theta) - \hat{D}(\hat{\theta}))^\top] &\leq c_1 \delta_n^2 \\ \mathbb{E} \sup_{\|\hat{\theta} - \theta_0\| \leq \delta_n} [(\hat{D}(\theta_0) - \hat{D}(\hat{\theta})) \hat{W}^{-1} (\hat{D}(\theta_0) - \hat{D}(\hat{\theta}))^\top] &\leq c_2 \delta_n^2 \\ \mathbb{E} \sup_{\|\theta - \theta_0\| \leq \delta_n} [(D(\theta) - D(\theta_0))^\top \hat{W}^{-1} (D(\theta) - D(\theta_0))^\top] &\leq c_3 \delta_n^2, \end{aligned}$$

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 $|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$, we have

$$\begin{aligned} |\hat{D}(\theta) - \hat{D}(\hat{\theta})| &= \left| \mathbb{P}_n \left[\sum_{t=0}^{T-1} \left\{ \max_a \theta^\top \varphi(S_{t+1}, a) - \max_a \hat{\theta}^\top \varphi(S_{t+1}, a) - (\theta - \hat{\theta})^\top \varphi(S_t, A_t) \right\} \varphi(S_t, A_t)^\top \right] \right| \\ &\leq \mathbb{P}_n \left[\sum_{t=0}^{T-1} \|\varphi(S_t, A_t)\| \left\{ \sum_a \|\varphi(S_{t+1}, a)\| \|\theta - \hat{\theta}\| + \|\varphi(S_t, A_t)\| \|\theta - \hat{\theta}\| \right\} \right]. \end{aligned}$$

Thus,

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

where

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

Therefore

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

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

Part II (Asymptotic Normality). Let $\theta = \theta_0 + \frac{b}{\sqrt{n}}$ and

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

Then,

$$\begin{aligned} \hat{D}(\theta_0 + b/\sqrt{n}) &= \mathbb{P}_n \left[\sum_t \left\{ R_{t+1} + \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) - \theta_0^\top \varphi(S_t, A_t) - b^\top / \sqrt{n} \varphi(S_t, A_t) \right. \right. \\ &\quad \left. \left. - \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right] \\ &= \mathbb{P}_n \left[\sum_t \left\{ \delta_{t+1} - b^\top / \sqrt{n} \varphi(S_t, A_t) - \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) \right. \right. \\ &\quad \left. \left. + \gamma \max_a (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right]. \end{aligned}$$

Using the above equation and the defined δ_{t+1} , the function $\hat{V}(b)$ can be written as

$$\begin{aligned} n\mathbb{P}_n \left[\sum_t \left\{ 2\delta_{t+1} - b^\top / \sqrt{n}\varphi(S_t, A_t) - \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right] \hat{W}^{-1} \\ \times \mathbb{P}_n \left[\sum_t \left\{ -b^\top / \sqrt{n}\varphi(S_t, A_t) - \gamma \max_a \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \right\} \varphi(S_t, A_t)^\top \right]^\top. \end{aligned}$$

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

- I. $n\mathbb{P}_n [\sum_t b^\top / \sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top] \hat{W}^{-1} \mathbb{P}_n [\sum_t b^\top / \sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top]^\top \rightarrow_p b^\top W b$
- II. $n\mathbb{P}_n [\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top] \hat{W}^{-1} \mathbb{P}_n [\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top]^\top \rightarrow_p \mathbb{E} [\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top] W^{-1} \times \mathbb{E} [\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top]^\top$
- III. $-2n\mathbb{P}_n [\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top] \hat{W}^{-1} \mathbb{P}_n [\sum_t b^\top / \sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top]^\top \rightarrow_d -2Z_\infty b$
- IV. $2n\mathbb{P}_n [\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top] \hat{W}^{-1} \mathbb{P}_n [\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top]^\top \rightarrow_d 2Z_\infty W^{-1} \mathbb{E} [\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top]^\top$
- V. $-2n\mathbb{P}_n [\sum_t b^\top / \sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top] \hat{W}^{-1} \mathbb{P}_n [\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top]^\top \rightarrow_p -2b^\top \mathbb{E} [\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top]^\top$

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

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

$$\zeta_{t+1}(\theta) = -\gamma \max_{a \in \pi^*} (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) + \gamma \max_a (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) + \gamma \max_{a \in \pi^*} b^\top / \sqrt{n}\varphi(S_{t+1}, a).$$

Thus, by Lemma 1, when $\|b/\sqrt{n}\| \rightarrow 0$ as $n \rightarrow \infty$, we have

$$\sqrt{n}[\zeta_{t+1}(\theta)] \rightarrow \psi_{t+1},$$

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

$$\sqrt{n}\mathbb{P}_n \left[\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top \right] \rightarrow_p \mathbb{E} \left[\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top \right]^\top.$$

We just showed that $\hat{V}(b)$ converges in distribution to

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

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

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

which is a consequence of the epi-convergence results of Geyer (1994). Note that when $\gamma = 0$, $\arg \min_b V(b) = Z_\infty W^{-1}$. Also, under assumption A.8, that is, when $|\pi^*| = 1$, and γ is small enough,

$\arg \min_b V(b) = Z_\infty \Gamma$, where

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

where I is an identity matrix. Hence,

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

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

S4. THE EFFECT OF THE CHOICE OF γ 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 γ . To better reflect the effect of γ , we assume that there is no death (i.e., $C_t=0$ for all $t = 0, \dots, 15$), and the reward function is defined as

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

For smaller values of γ ($\gamma = 0.1$), the estimated optimal policy will be more myopic and does not suggest augmenting any medication. This happens because the side effect outweighs the treatment effect. However, as γ gets larger, the optimal policy suggests to augment more treatments simply because the long-term effect of treatments outweighs the side effects. This shows that γ balances the immediate and long-term effect of treatments. Results are presented in Figure 5.

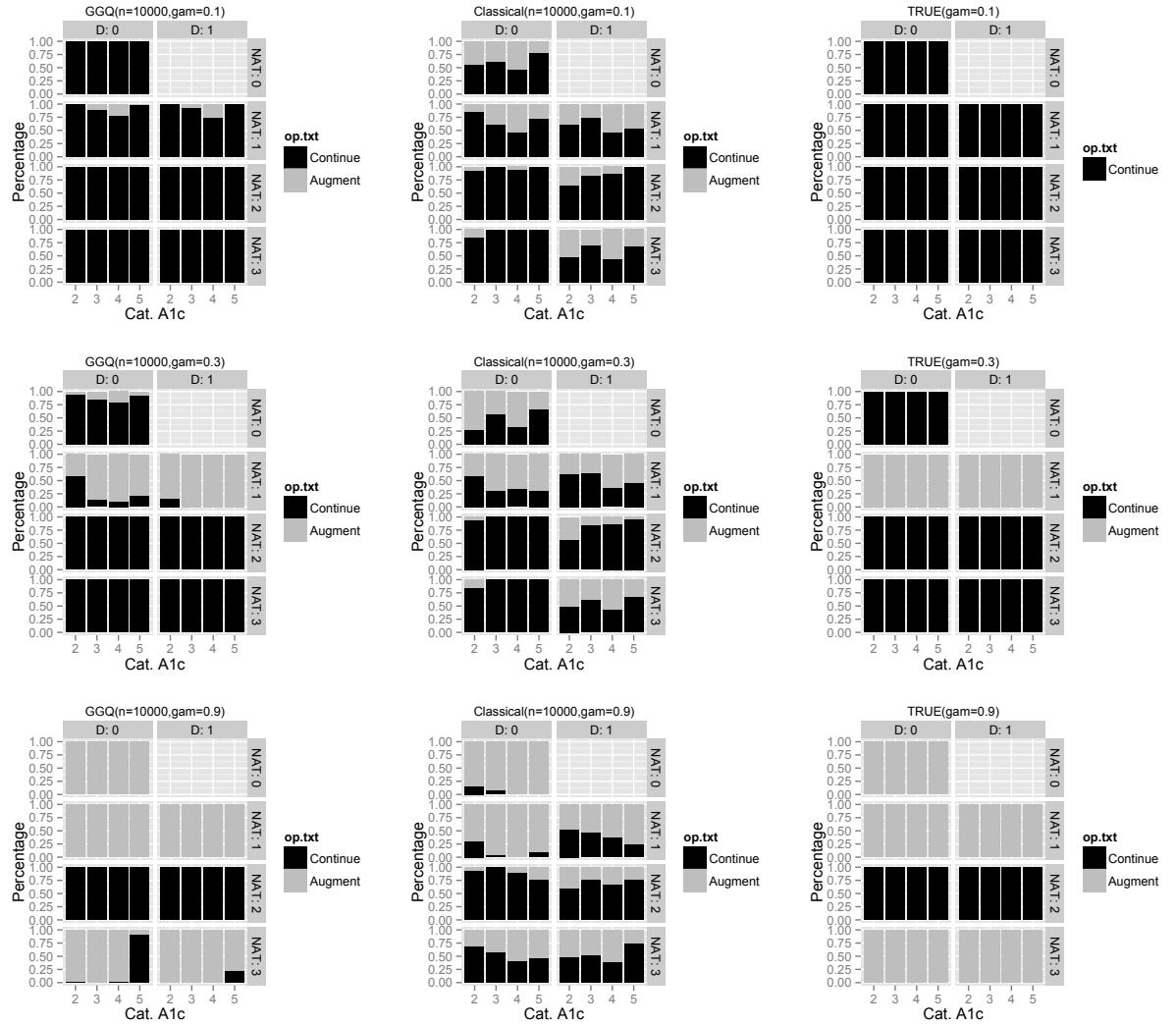
S5. THE EFFECT OF THE CHOICE OF TUNING PARAMETERS ON THE ESTIMATED OPTIMAL REGIME.

In this section, we discuss the effect of the tuning parameters (α, β) on the estimated optimal treatment regime under the simulation scenario discussed in the manuscript. We generated 500 datasets of size 2000 and applied the proposed methods using various choices of tuning parameters:

1. $\alpha = \frac{\nu}{k \log(k)}$ and $\beta = \frac{\nu}{k}$ for $\nu = 0.050, 0.025$, and 0.010
2. $\alpha = \frac{\nu}{k}$ and $\beta = \frac{\nu}{k^{3/4}}$ for $\nu = 0.050, 0.025$, and 0.010
3. $\alpha = \frac{\nu}{k}$ and $\beta = \frac{\nu}{k^{1/3}}$ for $\nu = 0.050$, and 0.010 .

The parameter ν specifies the step size (increment size) for each choice of the tuning parameters. Figure 6 shows that as long as the step sizes are not very small, the stochastic minimization algorithm has a good performance. However, when the step sizes are small ($\nu = 0.010$), the algorithm fails to converge to the true values because it cannot reach the true minimizers of the objective function. Table 1 presents the value of the objective function at the estimated $\hat{\theta}$ and the number of required iterations to converge (K) using different tuning parameters. The value of the objective function $M(\theta)$ for small ν is more than twice the value of $M(\theta)$ for larger values of ν , which indicates the lack of convergence to the true minimizers. Based on this result, the first choice of tuning parameter, $\nu = 0.05$, outperforms the other choices.

Figure 7 displays the effect of tuning parameters on standard errors. The vertical axis is the ratios of the standard errors obtained by different simulation scenarios over the standard error obtained by $\alpha = \frac{\nu}{k \log(k)}$

Fig. 5: Simulation: The effect of the choice of γ on the estimated optimal regime.

and $\beta = \frac{\nu}{k}$ and $\nu = 0.050$. For example, the vertical axis in the first plot is

$$\begin{aligned} & \text{S.D. of } \hat{\theta} \text{ when } \alpha = \frac{\nu}{k \log(k)}, \beta = \frac{\nu}{k} \text{ and } \nu = 0.025 \\ & \text{S.D. of } \hat{\theta} \text{ when } \alpha = \frac{\nu}{k \log(k)}, \beta = \frac{\nu}{k} \text{ and } \nu = 0.050. \end{aligned}$$

Note that the reference S.D.s in the denominator includes the tuning parameter values used in the main simulation study in Section 4, which is shown to converge to the true values. In the first two rows, the ratios deviate more from one as the step size (ν) gets smaller. This can be due to stopping the updates before converging to the minimizer of the objective function.

A. ERTEFAIE

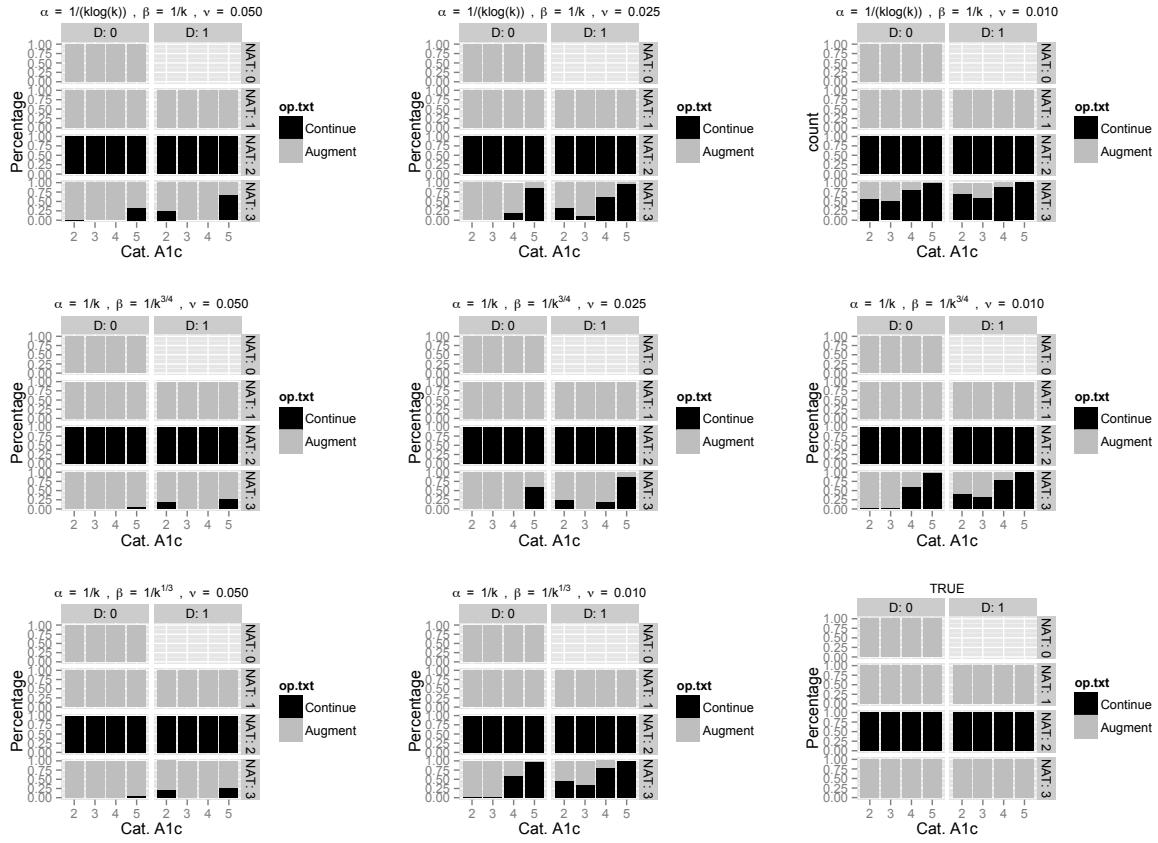


Fig. 6: Simulation: The effect of the choice of tuning parameters on the estimated optimal regime.

Table 1: Simulation: The effect of the choice of tuning parameters on the estimated optimal regime.

	$\alpha = \frac{1}{k \log(k)}$, $\beta = \frac{1}{k}$ $\nu = 0.05$ $\nu = 0.025$ $\nu = 0.01$			$\alpha = \frac{1}{k}$, $\beta = \frac{1}{k^{3/4}}$ $\nu = 0.05$ $\nu = 0.025$ $\nu = 0.01$			$\alpha = \frac{1}{k}$, $\beta = \frac{1}{k^{1/3}}$ $\nu = 0.05$ $\nu = 0.01$	
$M(\theta)$	0.007	0.011	0.025	0.007	0.008	0.017	0.007	0.018
K	14.03	13.22	12.86	21.35	20.66	19.44	21.77	19.45

REFERENCES

- BATHER, J. (2000). *Decision theory: an introduction to dynamic programming and sequential decisions*, vol. 180. Wiley Hoboken, NJ.
- BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. & WELLNER, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press.
- CHAKRABORTY, B., MURPHY, S. & STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19**, 317–343.
- COLLINS, L., MURPHY, S. & BIERMAN, K. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science* **5**, 185–196.
- GEYER, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics* **2**, 1993–2010.
- GOLDBERG, Y. & KOSOROK, M. R. (2011). Q-learning with censored data. *Annals of Statistics* **40**, 529–560.

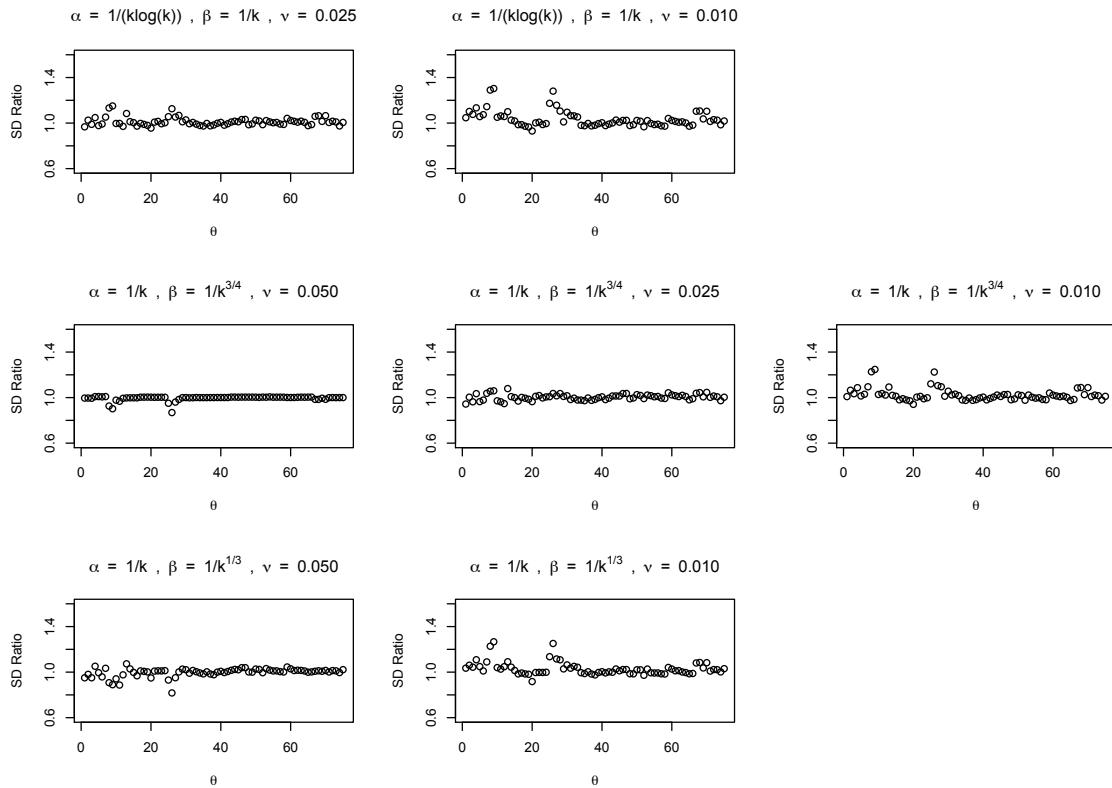


Fig. 7: Simulation: The effect of the choice of tuning parameters on the standard errors.

- 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**, 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**, 2427–2443.
- LABER, E., QIAN, M., LIZOTTE, D. J. & MURPHY, S. A. (2010). Statistical inference in dynamic treatment regimes. *arXiv preprint arXiv:1006.5831*.
- LAVORI, P. & 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**, 29–38.
- MAEI, H., SZEPESVÁRI, C., BHATNAGAR, S. & SUTTON, R. (2010). Toward off-policy learning control with function approximation. *Proc. ICML 2010*, 719–726.
- MANNOR, S., SIMESTER, D., SUN, P. & TSITSIKLIS, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science* **53**, 308–322.
- MOODIE, E., RICHARDSON, T. & STEPHENS, D. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63**, 447–455.
- MOODY, J. & DARKEN, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural computation* **1**, 281–294.
- MURPHY, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 331–355.
- MURPHY, S., OSLIN, D., RUSH, A. & ZHU, J. (2006). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* **32**, 257–262.

- NAHUM-SHANI, I., QIAN, M., ALMIRALL, D., PELHAM, W. E., GNAGY, B., FABIANO, G. A., WAXMONSKY, J. G., YU, J. & MURPHY, S. A. (2012). Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods* **17**, 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. & 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*.
- POGGIO, T. & GIROSI, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE* **78**, 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**, 1393–1512.
- ROBINS, J. (1987). Addendum to a new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Computers & Mathematics With Applications* **14**, 923–945.
- ROBINS, J. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*. Springer: New York.
- ROBINS, J., ORELLANA, L. & ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* **27**, 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**, 2379–2412.
- ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*. Springer, pp. 69–117.
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6**, 34–58.
- SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. & 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*, vol. 2. Wiley-IEEE Press.
- SIMESTER, D. I., SUN, P. & TSITSIKLIS, J. N. (2006). Dynamic catalog mailing policies. *Management Science* **52**, 683–696.
- SUTTON, R. & BARTO, A. (1998). *Reinforcement learning: An introduction*, vol. 28. Cambridge Univ Press.
- SUTTON, R., MAEI, H., PRECUP, D., BHATNAGAR, S., SILVER, D., SZEPESVÁRI, C. & 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*. ACM.
- SUTTON, R., SZEPESVÁRI, C. & MAEI, H. (2009b). A convergent $\alpha(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in Neural Information Processing Systems* **21**.
- TANG, Y. & KOSOROK, M. R. (2012). Developing adaptive personalized therapy for cystic fibrosis using reinforcement learning. Tech. rep., The University of North Carolina at Chapel Hill.
- TIMBIE, J., HAYWARD, R. & 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**, 437–456.
- VAN DER VAART, A. & WELLNER, J. (1996). *Weak convergence and empirical processes*. Springer Verlag.
- VAPNIK, V., GOLOWICH, S. E. & SMOLA, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, 281–287.
- ZHANG, B., TSIATIS, A. A., LABER, E. B. & DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.
- ZHANG, B., TSIATIS, A. A., LABER, E. B. & DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100**, 1–14.
- ZHAO, Y., KOSOROK, M. R. & ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine* **28**, 3294–3315.
- ZHAO, Y., ZENG, D., SOCINSKI, M. A. & KOSOROK, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–1433.