Estimating Optimal Dynamic Treatment Assignment Rules under Intertemporal Budget Constraints

Shosei Sakaguchi*

Preliminary draft

March, 2019

Abstract

This paper studies a statistical decision rule for the dynamic treatment assignment problem. Many policies involve dynamics in their treatment assignments, where treatments are sequentially assigned to individuals over multiple stages. In the dynamic treatment policies, the effect of each stage of treatment is usually heterogeneous depending on the past treatment assignments, associated outcomes, and observed covariates. We suppose that the policy maker wants to know the dynamic treatment assignment rule that guides the optimal treatment assignment at each stage based on the history of treatment assignments, outcomes, and observed covariates. This paper proposes the empirical welfare maximization method in the dynamic framework, which estimates the optimal dynamic treatment assignment rule from panel data of experimental or quasi-experimental studies. To solve the optimization problem that arises from the direct and indirect effect of each stage of treatment on future outcomes, I propose two estimation methods: one solves the whole dynamic treatment assignment problem simultaneously and the other solves each stage of the treatment assignment problem through backward induction. I derive uniform finite-sample bounds on the worst-case regret for the estimated rules and show $n^{-1/2}$ convergence rates. I also modify these estimation methods to incorporate intertemporal budget constraint, and provide finite-sample bounds for the regret and the deviation of the implementation cost of the estimated rule from the actual budget.

Keywords: Dynamic treatment effect, dynamic treatment regime, individualized treatment rule, empirical welfare maximization.

*Department of Economics, University College London, Gower Street, London WC1E 6BT, UK. E-mail: s.sakaguchi@ucl.ac.uk.

1 Introduction

Many policies involve dynamics in their treatment assignments. Some policies assign a series of treatments on individuals over multiple stages. For example, there are some job training programs that are composed of multiple stages and at each stage a different training is provided (e.g., Lechner, 2009; Rodríguez et al., 2018). Some other policies are characterized by different timings to initiate or terminate treatments. Important examples are unemployment insurance policies where one concern is the timing of reducing insurance (e.g., Meyer, 1995; Kolsrud et al. 2018). Aside from them, examples of dynamic treatment assignments include sequential medical interventions, educational interventions, or online advertisements. Because many treatment assignments involve dynamics, the dynamic treatment analysis has been attracting increasing attention (Abbring and Heckman, 2007).

For dynamic treatment assignment policies, policy makers want to know how to assign a series of treatments over stages depending on individuals' accumulated information at each stage in order to maximize social welfare. In the sequential job training programs, they want to know how to assign the series of trainings to each individual at each stage depending on his/her history of treatments, associated outcomes, and observed characteristics. In the unemployment insurance policies, an important question is when to reduce the insurance for each individual depending on his/her characteristics and past efforts on job finding.

This paper develops a statistical decision method to solve the dynamic treatment choice problems using panel data from experimental or quasi-experimental studies. We assume dynamic unconfoundedness (Robins, 1989, 1997) meaning that the treatment assignment at each stage is independent of current and future potential outcomes conditionally on observed characteristics and history of past treatment assignments and outcomes. Under this assumption, I construct the method to estimate the optimal Dynamic Treatment Regime (DTR)[1] by extending a method proposed by Kitagawa and Tetenov (2018a) into the dynamic treatment assignment framework. In the static framework, building on classification methods in machine learning, Kitagawa and Tetenov (2018a) develop the method, what they call Empirical Welfare Maximization (EWM) rule, to estimate optimal treatment assignment rules when exogenous constraints are placed on treatment assignment. The remarkable features of the EWM rule are its capabilities to accommodate exogenous policy constraints from legal, ethical, or political reasons and budget or capacity constraints and to restrict complexity of treatment assignment rules. The method I propose in this paper maintains these features. I call the proposed method Dynamic Empirical Welfare Maximization (DEWM) rule.

Further, as a specific feature of the DEWM rule, we can specify different types of dynamic treatment assignment problems: (i) sequential treatment assignment problems, where one of several

[1] Borrowing the terminology in statistics literatures, I call the dynamic treatment assignment rule DTR.

treatments is assigned at each stage and the analyst's goal is to make a dynamic protocol by which the policy maker can choose an optimal treatment for each individual at each stage depending on the individual's stage-specific information[2]; (ii) treatment timing problems, where the goal is to decide a rule by which the policy maker can decide to initiate or terminate a treatment at each stage depending on the accumulated information of the corresponding stage.[3] The DEWM rule can specify each type of these problems by constraining the class of feasible DTRs.

The dynamic treatment framework has several specific characteristics, which make it nontrivial to extend the original EWM rule to the dynamic setting. One is that the effect of treatment at each stage varies depending on the past treatment assignments and outcomes[4], so that the treatment at each stage should be decided by taking account of not only its direct effects on future outcomes but also its indirect effects through changing the effects of the future treatments. I solve this problem by providing two approaches. One approach is to estimate the optima DTR simultaneously, that is solving the whole sample welfare maximization problem with respect to the entire DTR simultaneously. The other approach is to estimate the optimal DTR through backward induction, where the treatment choice problem at each stage is solved from the final stage to the first stage supposing at each stage that the optimal treatments are chosen in the future stages. The second problem is that, in dynamic treatment policies, budget or capacity constraints are usually imposed intertemporary. Thus, the preferable treatment assignment rule should effectively allocate the intertemporal budget/capacity for each stage. This problem is solved by imposing the intertemporal budget/capacity constraints into the welfare maximization problem as optimization restrictions and estimates the treatmetn assignment rule that shoul satisfy the budget constraint.

I evaluate the statistical performance of the DEWM rule in terms of regret that is the average welfare loss relative to the maximum welfare achievable in a class of feasible DTRs. I derive finite-sample and distribution-free upper bounds on the regrets of the two methods in terms of the sample size n, a measure of complexity of the class of feasible DTRs, and the number of policy stages T. I show that these regrets converge to zero at rate $n^{-1/2}$. When the intertemporal budget/capacity constraints are imposed, I also analyze the deviation of the implementation cost of the estimated DTR from the actual budget/capacity constraints in terms of probability approximation.

This paper is related to the literature of treatment assignment rule, but most works in the literature focus on the static treatment assignment rule.[5][6] Han (2019) studies the identification

[2]Examples include the sequential job training programs (e.g., Lechner, 2009; Rodríguez et al., 2018).

[3]Examples include the unemployment insurance policies (e.g., Meyer, 1995; Kolsrud et al. 2018) and the work practice program for the unemployed (e.g., Vikström, 2017).

[4]In other words, the treatment at each stage influences on future outcomes not only through its direct effect but also through changing the effects of future treatments (indirect effects).

[5]A partial list of works in the literature are Manski (2004), Dehejia (2005), Hirano and Porter (2009), Chamberlain (2011), Bhattacharya and Dupas (2012), Stoye (2012), Tetenov (2012), Kasy (2014), Armstrong and Shen (2015), Athey and Wagner (2017), Kitagawa and Tetenov (2018a,c), Kock and Thyrsgaard (2018), and Mbakop and Tabord-Meehan (2018). Kitagawa and Tetenov (2018a) provides a detailed survey of these works.

[6]Note that the dynamic treatment framework in this paper is different from that in Kock and Thyrsgaard (2018).

of dynamic treatment effects and optimal DTRs relying on the instruments excluded from the outcome-determining process and other exogenous variables excluded from the treatment-selection process.[7] Although it resolves some issues of the dynamic unconfoundedness assumption such as noncompliance, the identified DTR is somewhat inflexible, in that it depends only on the pre-treatment covariates and cannot accommodate the exogenous constraints on assignment.

Estimation of optimal DTRs has been studied in medical statistics, under the labels of dynamic treatment regime, adaptive strategies or adaptive interventions, and various methods have been proposed.[8] A common approach is to estimate models for the conditional mean or other aspects of the conditional distributions of the outcomes and, then, solve the optimal DTR with approximating dynamic programming. This approach includes Q-learning (Murphy, 2005; Moodie, et al., 2012) and A-learning (Murphy, 2003; Robins, 2004) which, respectively, specify models of the stage-specific conditional mean outcome and regret with respect to current and history of treatments, outcomes, and covariates. A potential drawback of this approach is that the estimator of the optimal DTR requires the correctly specified outcome models even when using experimental data. Based on classification methods, Zhao et al. (2015) develops the estimation method of the DTR using a Support Vector Machine, which does not specify outcome models. They also derived the welfare convergence rates that depend on the sample size and the dimension of the accumulated information at each stage. Their approach is computationally attractive because of its use of a surrogate loss, but it cannot accommodate the exogenous constraints on assignment or the budget/capacity constraints.

The reminder of the paper is structured as follows. Section 2 describes the dynamic treatment framework, following Robins (1986; 1987), and defines the dynamic treatment assignment problem. Section 3 presents the two types of the DEWM methods and provides their statistical properties. In section 4, I modify one of the methods into the case that the intertemporal budget/capacity constraints are imposed. I conclude this paper in Section 5.

2 Setup

I first introduce the dynamic treatment framework, following Robins's counterfactual framework (Robins, 1986; 1987), in Section 2.1. Subsequently, in Section 2.2, I formalize the dynamic treatment assignment problem which a policy maker wants to solve.

Kock and Thyrsgaard (2018) consider the bandit problem setting where different individuals gradually come to each treatment assignment stage and do not receive multiple stages of treatments.

[7] Heckman and Navarro (2007) and Heckman et al. (2016) also study identification of dynamic treatment effects without relying on the dynamic unconfoundedness assumption.

[8] Chakraborty and Murphy (2014) review the developments in this field.

2.1 Dynamic Treatment Framework

We suppose that there are T ($T \geq 2$) stages of binary treatment assignment and, at each stage, an outcome is observed after the treatment is assigned. The treatments may be different across stages.[9] Let the binary treatment at each stage t be denoted by $D_t \in \{0,1\}$ for $t = 1, \ldots, T$. Throughout this paper, for any variable A_t, we denote by $\underline{\mathrm{A}}_t = (A_1, \ldots, A_t)$ a history of the variable up to stage t. The history of treatment assignments up to stage t is denoted by $\underline{\mathrm{D}}_t = (D_1, \ldots, D_t)$. Depending on the prior history of treatment assignments, we observe the outcome at each stage t which we denoted by $Y_t \in \mathbb{R}$. Let $Y_t(\underline{\mathrm{d}}_t)$ be a potential outcome at stage t that is realized when the history of treatment assignments up to stage t corresponds to $\underline{\mathrm{d}}_t \in \{0,1\}^t$. Then, the observed outcome at stage t is expressed as

$$Y_t = \sum_{\underline{\mathrm{d}}_t \in \{0,1\}^t} 1\{\underline{\mathrm{D}}_t = \underline{\mathrm{d}}_t\} Y_t(\underline{\mathrm{d}}_t),$$

where $1\{\cdot\}$ denotes the indicator function. Let X_t be k-dimensional covariates that is observed before a treatment is assigned at stage t. X_t may depend on the past treatment assignments and outcomes as well as their past values. For the first period, X_1 represents the pre-treatment information which contains individuals' demographic characteristics observed before the dynamic treatment policy starts. Let $H_t = \left(\underline{\mathrm{D}}_{t-1}, \underline{\mathrm{Y}}_{t-1}, \underline{\mathrm{X}}_t\right)$ denote the history of all the observed variables up to stage t, which is available information for the policy maker when choosing t-th stage of treatment. Note that $H_1 = (X_1)$. We denote the support of H_t by $\mathcal{H}_t$. Let P denote the distribution of $\left(D_t, \{Y_t(\underline{\mathrm{d}}_t)\}_{\underline{\mathrm{d}}_t \in \{0,1\}^t}, X_t\right)_{t=1}^T$.

From an experimental or quasi–experimental study, we observe $Z_i = (D_{it}, Y_{it}, X_{it})_{t=1}^T$ for individuals $i = 1, \ldots, n$ from the distribution of $(D_t, Y_t, X_t)_{t=1}^T$. Let $e_t(d_t, h_t) = \Pr(D_t = d_t \mid H_t = h_t)$ be a propensity score of treatment assignment at stage t given the history up to the corresponding stage. We suppose it is known to researcher under an experimental study, but it is unknown and need to be estimated under a observational study. We consider the case of the experimental study in this paper. The case of the observational study is ongoing work.

For further analysis, we suppose that the following assumptions hold.

Assumption 2.1. The vectors Z_i, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d).

Assumption 2.2. Sequential Independence Assumption: For each $t = 1, \ldots, T$ and $\underline{\mathrm{d}}_T \in \{0,1\}^T$, $D_t \perp (Y_t(\underline{\mathrm{d}}_t), \ldots, Y_T(\underline{\mathrm{d}}_T)) \mid H_t = h_t$ for any $h_t \in \mathcal{H}_t$.

[9] For example, in the two-stages of job training programs, the trainings may differ between the stages such that the second stage of training is more intensive than another.

Assumption 2.3. (i) Bounded Outcomes: There exists $M_t < \infty$ such that the support of the outcome variable Y_t is contained in $[-2/M_t, 2/M_t]$ for each $t = 1, \ldots, T$.
(ii) Overlap Condition: There exists $\kappa_t \in (0, 1/2)$ such that $e_t(1, h_t) \in [\kappa_t, 1 - \kappa_t]$ for all $h_t \in \mathcal{H}_t$ at each $t = 1, \ldots, T$.

The first assumption is a usual i.i.d assumption, where we do not impose any restriction on the distribution over the stages. The second assumption is what is called dynamic unconfoundedness assumption or sequential/dynamic conditional independence assumption, which is commonly used in the literature of the dynamic treatment regime (Robins 1986, 1987; Murphy, 2003; Lechner and Miquel, 2010). This assumption means that treatment assignment at each stage is independent of current and future potential outcomes conditionally on the past treatment assignments and the realized outcomes as well as covariates history. This assumption is usually satisfied under sequential randomization experiments. Under observational studies, this assumption is sometimes controversial but can hold if sufficient set of confounders and history of treatment assignments and outcomes are available (e.g., Lechner, 2009; Vikström, 2017). The third assumption is commonly assumed in the literature of the treatment effect analysis.

2.2 Dynamic Treatment Choice Problem

The goal of this paper is providing methods to estimate the optimal DTR form experimental or quasi-experimental panel data. We denote the treatment assignment rule at each stage t by $g_t : \mathcal{H}_t \mapsto \{0, 1\}$, that is a mapping from the history up to stage t to the treatment assignment of the corresponding stage, and define the DTR by $g = (g_1, \ldots, g_T) \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_T$, a sequence of the stage-specific treatment assignment rules. Thus, the DTR chooses treatment at each stage depending on the corresponding history.

We suppose that the welfare the policy maker wants to maximize is the population mean of the weighted sum of outcomes, $E_P\left[\sum_{t=1}^{T} \gamma_t Y_t\right]$, where the weight γ_t, for $t = 1, \ldots, T$, lies in $[0, 1]$ and is chosen by the policy maker. If the policy maker targets a discounted welfare, the weights are set to $\gamma_t = \gamma^{T-t-1}$, for $t = 1, \ldots, T$, where $\gamma \in (0, 1)$ is a discounted factor. If the policy maker targets the final outcome only, the weight are set to $\gamma_t = 0$ for $1 \leq t \leq T - 1$ and $\gamma_T = 1$.

Under a certain DTR g, the realized welfare takes the following form:

$$W(g) \equiv E_P\left[\sum_{\underline{d}_T \in \{0,1\}^T} \left(\prod_{t=1}^{T} 1\{g_t(H_t) = d_t\} \cdot \sum_{t=1}^{T} \gamma_t Y_t(\underline{d}_t)\right)\right].$$

Under Assumption 2.2, given the propensity score $e_t(d_t, h_t)$, the welfare can be written equiv-

alently as

$$W(g) = E_P\left[\sum_{\underline{d}_T \in \{0,1\}^T} \frac{\left(\sum_{t=1}^T \gamma_t Y_t(\underline{d}_t)\right) \cdot 1\{\underline{D}_T = \underline{d}_T\} \cdot \prod_{t=1}^T 1\{g_t(H_t) = d_t\}}{\prod_{t=1}^T e_t(d_t, H_t)}\right] \tag{2.1}$$

$$= E_P\left[\sum_{d_1 \in \{0,1\}} \frac{\gamma_1 Y_1 \cdot 1\{\underline{D}_1 = \underline{d}_1\} \cdot \prod_{t=1}^1 1\{g_t(H_t) = d_1\}}{\prod_{t=1}^1 e_t(d_t, H_t)}\right]$$

$$+ \cdots + E_P\left[\sum_{\underline{d}_T \in \{0,1\}^T} \frac{\gamma_T Y_T \cdot 1\{\underline{D}_T = \underline{d}_T\} \cdot \prod_{t=1}^T 1\{g_t(H_t) = d_t\}}{\prod_{t=1}^T e_t(d_t, H_t)}\right]. \tag{2.2}$$

In this paper, following Kitagawa and Tetenov (2018a), we restrict the complexity of the class of feasible DTRs in terms of VC-dimension. We denote the class of feasible DTRs by $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_T$, where $\mathcal{G}_t$ is a class of t-th stage of treatment assignment rule g_t. We impose the following assumption.

Assumption 2.4. VC-class: For each $t = 1, \ldots, T$, a class of function $\mathcal{G}_t$ of g_t is a VC-class of function and has VC-dimension $v_t < \infty$.

This assumption restricts the complexity of the class of whole DTRs through restricting the complexity of each class of stage-specific treatment assignment rule. By restricting the complexity, we can choose a simply explainable DTR, which is favorable for the policy maker. Some examples of the practically relevant classes of feasible DTRs are Linear Eligibility Score rule and Threshold Allocation rule.

Example 2.1. The class of DTRs based on the Linear Eligibility Score is $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_t$ where $\mathcal{G}_t$ for each $t \in \{1, \ldots, T\}$ is the following:

$$\mathcal{G}_t = \Big\{1\left\{\beta_1' \underline{x}_{t-1} + \beta_2' \underline{d}_{t-1} + \beta_3'\left(\left(1 - \underline{d}_{t-1}\right) \circ \underline{y}_{t-1}\right) + \beta_4'\left(\underline{d}_{t-1} \circ \underline{y}_{t-1}\right) \geq c\right\} :$$
$$: (\beta_1, \beta_2, \beta_3, \beta_4, c) \in \mathbb{R}^{k \times (t-1)} \times \mathbb{R}^{(t-1)} \times \mathbb{R}^{(t-1)} \times \mathbb{R}^{(t-1)} \times \mathbb{R}\Big\}.$$

Under this class of DTRs, treatment assignment at each stage is decided based on whether the eligibility score exceeds a certain threshold c. Note that, in the above specification, interactions terms of the past treatments and the corresponding stages of outcomes contribute to the eligibility scores, which means that, depending on the histoty of treatments, the edibility score evaluates the past outcomes differently. The main objective of data analysis under this class is to construct an eligibility score that presents a welfare-maximizing DTR. In this example, each $\mathcal{G}_t$ has VC-dimension $3k(t-1)+1$; thus, $\mathcal{G}$ has VC-dimension at most $(3k+1)T - 3k$.

Example 2.2. The class of DTRs based on the Threshold Allocation rule is the following: $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_t$ where $\mathcal{G}_t$ for each $t \in \{1, \ldots, T\}$ is

$$\mathcal{G}_t = \Big\{ 1\left\{ s_1 \circ \bar{x}_{t-1} \geq c_1, s_2 \circ \bar{d}_{t-1} \geq c_2, s_3 \circ \left[\left(1 - \bar{d}_{t-1}\right) \circ \bar{y}_{t-1} \right] \geq a_3, s_4 \circ \left(\bar{d}_{t-1} \circ \bar{y}_{t-1} \right) \geq a_4 \right\} \\ : (s_1, s_2, s_3, s_4) \in \{-1, 1\}^{k(t-1)+3(t-1)}, (c_1, c_2, c_3, c_4) \in \mathbb{R}^{k \times (t-1)} \times \mathbb{R}^{3(t-1)} \Big\}.$$

Under this class of DTRs, treatment is assigned at each stage if past covariates, treatment assignments, and realized outcomes exceed or fail certain thresholds. Then, what the data analyst does is to estimate the signs of $s_1, \ldots, s_4$ and the values of thresholds $c_1, \ldots, c_4$ so that the data-driven DTR maximizes the social welfare. In this example, each $\mathcal{G}_t$ has VC-dimension at most $3k\,(t-1)$ and, thus, VC-dimension of the whole class $\mathcal{G}$ is not more than $3k\,(T-1)$.

Aside from the restriction on the class of each stage of treatment assignment rule, by restricting the intertemporal relationship among treatment assignments, we can specify each type of dynamic treatment choice problem. We denote the restriction on whole class of g by $\tilde{\mathcal{G}}$. If the policy maker wants to decide a timing to assign a treatment that can be assigned only once for each individual, we should set $\tilde{\mathcal{G}} = \left\{ (g_1, \cdots, g_T) : \sum_{t=1}^{T} g_t = 1 \right\}$. If the problem is deciding a timing to initiate or terminate continuous treatment, we should set $\tilde{\mathcal{G}} = \{(g_1, \cdots, g_T) : g_s \leq g_t \text{ for } s \leq t\}$ or $\tilde{\mathcal{G}} = \{(g_1, \cdots, g_T) : g_s \leq g_t \text{ for } s \geq t\}$, respectively. Further, we can treat the problem of choosing both timings to initiate and terminate continuous treatment by setting

$$\tilde{\mathcal{G}} = \{(g_1, \cdots, g_T) : \text{if } g_j = 0 \text{ for any } j \leq s,\ g_s \leq g_t \text{ for } t \geq s; \text{ otherwise } g_s \geq g_t \text{ for } t \geq s\}.$$

We can impose each restriction on the DTRs by redefining $\mathcal{G} = \left(\prod_{t=1}^{T} \mathcal{G}_t \right) \bigcap \tilde{\mathcal{G}}$. Note that VC-dimension of this class is not more than that of the original class $\prod_{t=1}^{T} \mathcal{G}_t$.

In the setting described above, we denote the highest social welfare that is attainable under the feasible DTR $\mathcal{G}$ by

$$W_{\mathcal{G}}^{*} = \max_{g \in \mathcal{G}} W(g). \tag{2.3}$$

We assume that the planner's goal is to estimate the optimal DTR in $\mathcal{G}$, that maximize the social welfare, from the sample $Z_1, \ldots, Z_n$. As in Kitagawa and Tetenov (2018a), we do not require the first best DTR[10] to be achievable in $\mathcal{G}$. In the following section, I provide two methods to estimate the optimal DTR and evaluate their statistical properties in terms of the maximum regret of the welfare.

[10] The first best DTR is a welfare-maximizing DTR that is achievable in the class of whole measurable DTRs.

3 Dynamic Empirical Welfare Maximization (DEWM)

In this section, I propose two DEWM methods. One is based on backward induction (dynamic programming) to solve the sequential treatment choice problem; the other is based on the simultaneous optimization of $W(g)$ with respect to g. After that, I evaluate the statistical properties of the two methods in terms of the maximum regret of the social welfare fucntion.

3.1 Backward Dynamic Empirical Welfare Maximization

We now suppose that generative distribution function P is known. In this case, we can solve the dynamic treatment assignment problem through dynamic programming (backward induction). Firstly, for the final stage T, we can obtain

$$g_T^* \in \arg\max_{g_T \in \mathcal{G}_T} Q_T(h_T, g_T),$$

where $Q_T(h_T, g_T) = E(\gamma_T Y_T \mid H_T = h_T, D_T = g_T(h_t))$. Here, $g_T^* : \mathcal{H}_T \to \{0,1\}$ is an optimal treatment assignment rule at the final stage leading a best treatment that maximizes the social welfare with respect to any prior history $h_T \in \mathcal{H}_T$. Recursively, from $t = T-1$ to $t = 1$, we can solve

$$g_t^* \in \arg\max_{g_t \in \mathcal{G}_t} Q_t(h_t, g_t),$$

where

$$\begin{aligned} Q_t(h_t, g_t) &= E\left[\gamma_t Y_t + \sum_{s=t+1}^{T} \max_{g_s \in \mathcal{G}_s} Q_s(H_s, g_s) \mid H_t = h_t, D_t = g_t(g_t)\right] \\ &= E\left[\gamma_t Y_t + \sum_{s=t+1}^{T} Q_s(H_s, g_s^*) \mid H_t = h_t, D_t = g_t(g_t)\right]. \end{aligned}$$

For any $t = 1, \ldots, T-1$,$Q_t(h_t, g_t)$ is the expected welfare that is achieved when the policy maker assigns treatment g_t at stage t and the optimal treatment are assigned in the future stages. In this procedure, we obtain the optimal treatment at each stage through the welfare maximization problem given that we know the optimal treatment assignments in the future stages. Thus, the whole sequence $g^* = (g_1^*, \ldots, g_T^*)$ corresponds to the solution of the whole welfare maximization problem (2.3).[11] Note here that, given the propensity scores, the expected welfare $Q_t(h_t, g_t)$, for

[11] This idea is what the Q-learning is based on (Murphy, 2005; Moodie, et al., 2012). The Q-learning is an approximate dynamic programming algorithm that uses regression models to estimate the Q-functions $Q_t(h_t, g_t)$ $t = 1, \ldots, T$. Linear models are typically used to approximate the Q-function.

$t = 1, \ldots, T$, can be written equivalently as

$$Q_t\left(h_t, g_t\right) = E\left[q_t\left(h_t, g_t; g_{t+1}^*, \ldots, g_T^*\right)\right],$$

where

$$q_t\left(h_t, g_t; g_{t+1}, \ldots, g_T\right) \equiv \sum_{(d_t,\ldots,d_T)\in\{0,1\}^{T-t+1}} \left(\prod_{s=t+1}^{T} 1\left\{g_s\left(H_{is}\right) = d_s\right\}\right) \times \left\{\frac{\left(\prod_{s=t}^{T} 1\left\{D_{is} = d_s\right\}\right) 1\left\{g_t\left(H_{it}\right) = d_t\right\}\left(\sum_{s=t}^{T}\gamma_s Y_{is}\right)}{\prod_{s=t}^{T} e_s\left(d_s, \mathrm{H}_{is}\right)}\right\}.$$

The first estimation method I propose is based on the sample analogue of the above backward induction procedure. I call this method Backward DEWM method. The Backward DEWM method first estimates $\hat{g}_T^B$ such that

$$\hat{g}_T^B \in \arg\max_{g_T\in\mathcal{G}_T} \frac{1}{n}\sum_{i=1}^{n} q_T\left(H_{it}, g_T\right).$$

Then, recursively, from $t = T-1$ to $t = 1$, it estimates $\hat{g}_t^B$ such that

$$\hat{g}_t^B \in \arg\max_{g_t\in\mathcal{G}_t} \frac{1}{n}\sum_{i=1}^{n} q_t\left(H_{it}, g_t; \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right),$$

where $\hat{g}_{t+1}^B, \ldots, \hat{g}_T^B$ are estimated prior to stage t. Note that each maximization can be carried with the same algorithm in the first step, but the weights for the weighted outcomes $\left(\sum_{s=t}^{T}\gamma_s Y_{is}\right)$ are different among stages. We denote the DTR obtained through the above procedure by $\hat{g}^B = \left(\hat{g}_1^B, \ldots, \hat{g}_T^B\right)$.

3.2 Simultaneous Dynamic Empirical Welfare Maximization

The second approach I propose is a sample analogue of the simultaneous maximization problem (2.3). Instead of maximizing the sample analogue of (2.1), we consider to maximize the sample analogue of (2.2), because that provides better non-asymptotic properties. We call the method Simultaneous DEWM method. Formally, the Simultaneous DEWM method estimates $\hat{g}^S = \left(\hat{g}_1^S, \ldots, \hat{g}_T^S\right)$ simultaneously such that

$$\left(\hat{g}_1^S, \ldots, \hat{g}_T^S\right) \in \arg\max_{g\in\mathcal{G}} \sum_{t=1}^{T}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{\underline{\mathrm{d}}_t\in\{0,1\}^t} w_t^S\left(\underline{\mathrm{g}}_t, Y_{it}, D_{it}, H_{it}\right)\right],$$

where

$$w_t^S\left(\underline{g}_t, Y_{it}, D_{it}, H_{it}\right) = \frac{1\left\{\underline{D}_{it} = \underline{d}_t\right\} \cdot \left(\prod_{s=1}^t 1\left\{g_s\left(H_{is}\right) = d_s\right\}\right) \cdot \gamma_t Y_{it}}{\prod_{s=1}^t e_s\left(d_s, H_{is}\right)}.$$

Here, $n^{-1}\sum_{i=1}^n \sum_{\underline{d}_t \in \{0,1\}^t} w_t^S\left(\underline{g}_t, Y_{it}, D_{it}, H_{it}\right)$ corresponds to the sample analogue of the t-th term in (2.2).

Comparing between the two estimation methods, the Backward DEWM method is computationally attractive since it divides the maximization problem into T easier problems. However, when the intertemporal budget/capacity constraints are accomodated, the Simultaneous DEWM sometimes more computatinally attractive. We see this in more detail in the following section.

3.3 Statistical Properties

As in much of the literature that follows work of Manski (2004), we evaluate the statistical properties of the two DWEM methods, $\hat{g}^B$ and $\hat{g}^S$, in terms of the maximum regret relative to the optimal maximum feasible welfare $W_{\mathcal{G}}^*$. Following Kitagawa and Tetenov (2018a), we focus on the non-asymptotic upper bound of the worst-case average welfare loss $\sup_{p \in \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W\left(\hat{g}\right)\right]$, where $\mathcal{P}\left(M,k\right)$ is the class of distribution functions that satisfy Assumptions 2.1-2.3. The analysis refers to theoretical results established in classification literatures (e.g., Devroye et al., 1996; Mohry, 2008).

The following theorem provides a finite-sample upper bound on the average welfare loss and reveals its dependence on sample size n, VC-dimension v, and the number of stages T.

Theorem 3.1 Suppose Assumptions 2.1-2.4 hold. For any $j \in \{B, S\}$, we have

$$\sup_{p \in \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W\left(\hat{g}^j\right)\right] \leq 2C_1 \sum_{t=1}^T \left\{ \frac{\gamma_t M_t}{\prod_{s=1}^t \kappa_s} \sqrt{\frac{\sum_{s=1}^t v_s}{n}} \right\},$$

where C_1 is a some universal constant.

This theorem shows that the convergence rate of the worst-case welfare loss for the two DEWM rules is no slower than $n^{-1/2}$. The upper bound is increasing in the VC-dimension of $\mathcal{G}$, implying that, as the candidate treatment assignment rules become more complex in terms of VC-dimension, $\hat{g}$ tends to overfit the data in the sense that the distribution of regret is more and more dispersed.

The following proposition provides a different view for the worst-case welfare regret.

Proposition 3.1. Suppose Assumptions 2.1-2.4 hold. For $j \in \{B, S\}$ and any $\delta \in (0,1)$, the following holds with probability greater than $1-\delta$,

$$\sup_{p \in \mathcal{P}(M,\kappa)} \left| W_{\mathcal{G}}^{*} - W\left(\hat{g}^{j}\right) \right| \leq \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \left\{ \sqrt{8 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{2 \log\left(1/\delta\right)} \right\} \right] / \sqrt{n}.$$

This proposition provides the finite-sample upper bounds for the actual regret, rather than the average regret, that holds with high probability and also provides the guide to the choise of the sample size.

4 Budget/Capacity Constraints

In this section, we consider the budget/capacity constraints that restrict the proportion of the population that could be assigned to treatment. In the dynamic treatment policy, there should be two types of budget/capacity constraints: temporal and intertemporal budget/capacity constraints. The temporal budget/capacity constraints are imposed on each stage of treatment assignment independently and restrict the proportion of the population to be treated at each stage. The intertemporal constraints are simultaneously imposed on whole or multiple stages of treatment assignment. If there is a limited amounted of treatment or limited budget that can be expended at some specific-sate, this is the case when the policy maker faces a temporal budget/capacity constraint. On the other hand, if the policy maker has a budget that can be arbitrarily expended for multiple stages or limited amount of treatment can be assigned at any stage, this is the cases when an intertemporal budget/capacity constraint exists. I formalize these constraints in the following.

We suppose that the policy maker faces the following B constraints:

$$\sum_{t=1}^{T} K_{tb} E\left[D_t\right] \leq C_b \text{ for } b = 1, \ldots, B, \tag{4.1}$$

where $K_{tb} \in [0,1]$ and $C_b \geq 0$. For a scale normalization, we assume that $\sum_{t=1}^{T} K_{tb} = 1$ for all $b = 1, \ldots, B$. Here, for any $b = 1, \ldots, B$, the weights $K_{1b}, \ldots, K_{Tb}$ represent relative costs among stages of treatments and C_b represents the total capacity or budget of the policy. If $K_{tb} > 0$ and $K_{sb} = 0$ for any $s \neq t$, the b-th constraint corresponds to the temporal budget/capacity constraint for stage t. Otherwise, if at least two of $K_{1b'}, \cdots, K_{Tb'}$ take non-zero values, we regard theb'-th constraint as the intertemporal budget/capacity constraint. Especially, if all of $K_{1b'}, \cdots, K_{Tb'}$ take non-zero values, this is a budget/capacity constraint on the whole sequence of treatments. Note that B constraints may contain both the temporal and intertemporal constraints.

We suppose that the policy maker wants to maximize the social welfare under the budget/capacity constraints. For a feasible DTR class $\mathcal{G}$, the maximized social welfare is

$$W_{\mathcal{G}}^{*} = \sum_{t=1}^{T} \max_{g \in \mathcal{G}} W(g) \tag{4.2}$$
$$\text{subject to } \sum_{t=1}^{T} K_{tb} E[g_t] \leq C_b \text{ for } b = 1, \ldots, B.$$

The goal of the analysis is then to choose a DTR from $\mathcal{G}$ that achieves the maximized social welfare and satisfies the budget/capacity constraints.

To this end, I incorporate sample analogues of the budget/capacity constraints (3.1) into the Baskward and Simultaneous DEWM methods. The modified Simultaneous DEWM method then solves the following problem:

$$\left(\hat{g}_1^S, \ldots, \hat{g}_T^S\right) \in \arg\max_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\bar{d}_t \in \{0,1\}^t} w_t^S\left(\mathrm{g}_t, Y_{it}, D_{it}, H_{it}\right) \tag{4.3}$$

$$\text{subject to} \sum_{t=1}^{T} \left(K_{tb} \frac{1}{n} \sum_{i=1}^{n} g\left(H_{is}\right) \right) \leq C_b + \alpha_n \text{ for } b = 1, \ldots, B. \tag{4.4}$$

Here α_n is a tunable hyperparameter which takes positive value, depends on the sample size n and VC-dimension of $\mathcal{G}$, and converges to zero as n becomes large. This parameter is needed to makes the optimal DTR that solves (4.2) exists in the class of DTR that satisfy the sample budget/capacity constraints (4.4).

The following theorem shows finite-sample properties of the worst-case welfare loss of the modified Simultaneous DEWM method and further shows the deviation between the implementation costs of the optimal DTR and the estimated DTR that holds with high probability.

Theorem 4.1 Suppose Assumptions 2.1-2.4 hold. Let $W_{\mathcal{G}}^{*}$ be defined in (4.2) and $\hat{g}^S$ be a solution of (4.3). Then, for any $\delta \in (0,1)$, if $\alpha_n > \sqrt{\log(6B/\delta)/(2n)} \left(\max_{b \in \{1,\ldots,B\}} \sum_{t=1}^{T} K_{tb} \right)$, the following hold with probability greater than $1-\delta$:

$$\begin{aligned} &\sup_{p \in \mathcal{P}(M,\kappa)} \left| W_{\mathcal{G}}^{*} - W\left(\hat{g}^S\right) \right| \\ \leq & 2 \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log\left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6/\delta)}{2}} \right\} \right] / \sqrt{n} \end{aligned}$$

and

$$\sup_{p\in\mathcal{P}(M,\kappa)}\max_{b\in\{1,\ldots,B\}}\left(E_P\left[\sum_{t=1}^{T}K_{tb}\hat{g}^S\left(H_{it}\right)\right]-C_b\right) \tag{4.5}$$
$$\leq\alpha_n+2\sum_{t=1}^{T}\left[K_{tb}\left\{\sqrt{2\left(\sum_{s=1}^{t}v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t}v_s}\right)}+\sqrt{\frac{\log\left(6B/\delta\right)}{2}}\right\}\right]/\sqrt{n}.$$

Here, (4.5) means the deviation of the implementation costs of the estimated DTR from the actual budgets/capacities. The theorem shows that, if the sample size is large, the regret and the budget deviation is small. The worst-case welfare loss and the budget/capacity deviation diminish at rate $\sqrt{\left(\log n\right)/n}$.

If we consider the strict budget/capacity constraints:

$$\sum_{t=1}^{T}\left(K_{tb}\frac{1}{n}\sum_{i=1}^{n}g\left(H_{is}\right)\right)\leq C_b \text{ for } b=1,\ldots,B,$$

we have the following results with probability greater than $1-\delta$:

$$\sup_{p\in\mathcal{P}(M,\kappa)}\left|W_{\mathcal{G}}^{*}-W\left(\tilde{g}^S\right)\right|$$
$$\leq\sup_{p\in\mathcal{P}(M,\kappa)}\left|W_{\mathcal{G}}^{*}-W_{\mathcal{G}}^{\dagger}\right|+2\sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\left\{\sqrt{2\left(\sum_{s=1}^{t}v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t}v_s}\right)}+\sqrt{\frac{\log\left(6/\delta\right)}{2}}\right\}\right]/\sqrt{n}$$

and

$$\sup_{p\in\mathcal{P}(M,\kappa)}\max_{b\in\{1,\ldots,B\}}\left(E_P\left[\sum_{t=1}^{T}K_{tb}\tilde{g}^S\left(H_{it}\right)\right]-C_b\right)$$
$$\leq 2\sum_{t=1}^{T}\left[K_{tb}\left\{\sqrt{2\left(\sum_{s=1}^{t}v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t}v_s}\right)}+\sqrt{\frac{\log\left(6B/\delta\right)}{2}}\right\}\right]/\sqrt{n},$$

where $W_{\mathcal{G}}^{\dagger}$ it the optimal welfare with the constraints

$$\sum_{t=1}^{T}K_{tb}E\left[D_t\right]\leq C_b-\sqrt{\log\left(6B/\delta\right)/\left(2n\right)}\left(\max_{b\in\{1,\ldots,B\}}\sum_{t=1}^{T}K_{tb}\right) \text{ for } b=1,\ldots,B$$

and $g^{\dagger}=\left(g_1^{\dagger},\ldots,g_T^{\dagger}\right)$ is the associated optimal DTR. Note that $W_{\mathcal{G}}^{\dagger}$ the optimal welfare under

the budget that is smaller than the original budget. Here, $W_{\mathcal{G}}^{*} - W_{\mathcal{G}}^{\dagger}$ expresses the deviation of the optimal welfare with respect to the change of the budget constraint.

Next, we consider to incorporate the intertemporal budget/capacity constraints into the Backward DEWM method. Since the Backward DEWM method sequentially solves the each stage of the welfare maximization problem, we cannot incorporate the intertemporal constraints directry. Instead, we consider to seek the optimal allocation of the intertemporal budget/constraints to each stage of treatment assignment. Let $L = (L_1, \ldots, L_T)$ be the series of each stage of budget constraint that satifies

$$\sum_{t=1}^{T} K_{tb} L_t \leq C_b \tag{4.6}$$

for $b = 1, \ldots, B$. Further, define by $\hat{g}^B(L) = \left(\hat{g}_1^B(L_1), \ldots, \hat{g}_T^B(L_T)\right)$ the estimated DTR with Backward DWEM method under the constraints

$$K_{tb} \frac{1}{n} \sum_{i=1}^{n} g_t(H_{it}) \leq L_t$$

for any $b = 1, \ldots, B$ and $t = 1, \ldots, T$. We solve the welfare maximization problem with respect to not only g but also L, and denote the associated estimated rule and budget allocation, respectively, by $\hat{g}^B$ and $\hat{L}$. As in the case with the Simultaneous DEWM method, we need to modify the constraints (4.6) as follows:

$$\sum_{t=1}^{T} K_{tb} L_t \leq C_b + \alpha_n \text{ for } b = 1, \ldots, B,$$

where α_n is a tunable hyperparameter which takes positive value, depends on the sample size n and VC-dimension of $\mathcal{G}$, and converges to zero as n becomes large. This modification is needed to ensure that the optimal DTR g^* exsists in the class of the dynamic treatment regime that satisfies the sample budget/capacity constraints.

For the modifed Backward DEWM method, we have the following result.

Theorem 4.2 Suppose Assumptions 2.1-2.4 hold. Let $W_{\mathcal{G}}^{*}$ be defined in (4.2) and $\hat{g}^B$ be defined above. Then, for any $\delta \in (0,1)$, if $\alpha_n > \sqrt{\log(6B/\delta)/(2n)} \left(\max_{b \in \{1,\ldots,B\}} \sum_{t=1}^{T} K_{tb}\right)$, the following

hold with probability greater than $1-\delta$:

$$\begin{aligned}
&\sup_{p\in\mathcal{P}(M,\kappa)} \left|W_{\mathcal{G}}^{*} - W\left(\hat{g}^{B}\right)\right| \\
\leq & 2\sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\left\{\sqrt{2\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)} + \sqrt{\frac{\log\left(6/\delta\right)}{2}}\right\}\right]/\sqrt{n}
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{p\in\mathcal{P}(M,\kappa)} \max_{b\in\{1,\ldots,B\}}\left(E_P\left[\sum_{t=1}^{T} K_{tb}\hat{g}^{B}\left(H_{it}\right)\right] - C_b\right) \\
\leq & \alpha_n + 2\sum_{t=1}^{T}\left[K_{tb}\left\{\sqrt{2\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)} + \sqrt{\frac{\log\left(6B/\delta\right)}{2}}\right\}\right]/\sqrt{n}.
\end{aligned}$$

Here, the same argument to Theorem 4.1 is applied. Under the hard budget constraint, the result of Corrolarry 4.1 also holds for the Backward DEWM method.

5 Conclusion

In this paper, I propose empirical methods to estimate the the optimal DTR based on the empirical welfare maximization approach. The method can accommodate exogenous constraints on feasible DTRs and further specify the type of dynamic treatment choice problem through restricting the intertemporal relationship among multiple stages of treatments. I propose two estimation methods, the Simultaneous DEWM method and the Backward DEWM method, which estimate the optimal DTR, respectively, through simultaneous maximization and backward induction. I evaluate the finite-sample properties of these methods in terms of the worst-case welfare loss and derive theier uppoer bounds. These bounds show $n^{-1/2}$ convergence rates of the worst-case average welfare-loss towards zero for both the medhods. I further modify the Simultaneous DEWM method to incorporate the intertemporal budget/capacity constraints. I derive the finite-sample bounds of the actual worsta-case welfare loss and the deviation between the budget and the implementation cost of the estimated rule. The results show the consistency of the welfare loss and the budget constraint.

Appendix A.

This appendix provides the proofs of Theorems 3.1 and Proposition 3.1. Many concepts and techniques in the proofs owe to the literatures of classification (e.g., Devroye et al. 2009; Mohri et al. 2012). I first introduce the following lemma which will be used in the proof of Theorem 3.1.

Lemma A.1. (Kitagawa and Tetenov, 2018b, Lemma A.4) Let $\mathcal{F}$ be a class of uniformly bounded functions, that is, there exists $\bar{F} < \infty$ such that $\|f\|_\infty \leq \bar{F}$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph with VC-dimension $v < \infty$. Then, there is a universal constant C_1 such that

$$E_{P^n}\left[\sup_{f\in\mathcal{F}} |E_n(f) - E_P(f)|\right] \leq C_1 \bar{F}\sqrt{\frac{v}{n}} \tag{A.1}$$

holds for all $n \geq 1$.

The proof of this lemma is provided in Kitagawa and Tetenov (2018b). The following is the proof of Theorem 3.1.

(Proof of Theorem 3.1.)
I divide the proof in two parts. In the first part of the proof, I show the statement of the theorem hold for the Simultaneous DEWM method. The proof of the Backward DEWM method is provided in the second part.
(i) For the Simultaneous DEWM method:
Define

$$W_t(\mathrm{g}_t) \equiv E_P\left[\sum_{\underline{\mathrm{d}}_t\in\{0,1\}^t} \frac{\gamma_t Y_t \cdot 1\{\underline{\mathrm{D}}_t = \underline{\mathrm{d}}_t\} \cdot \prod_{s=1}^t 1\{g_s(H_s) = d_s\}}{\prod_{s=1}^t e_s(d_s, H_s)}\right],$$
$$w_t(\mathrm{g}_t, H_t) \equiv \sum_{\underline{\mathrm{d}}_t\in\{0,1\}^t} \frac{\gamma_t Y_t \cdot 1\{\underline{\mathrm{D}}_t = \underline{\mathrm{d}}_t\} \cdot \prod_{s=1}^t 1\{g_s(H_s) = d_s\}}{\prod_{s=1}^t e_s(d_s, H_s)},$$

and

$$w(g, Z) \equiv \sum_{t=1}^T w_t(\mathrm{g}_t, H_t).$$

Note that $W(g) = E[w(g, Z)]$ holds from the equation (2.2). Let $W_{nt}(\mathrm{g}_t)$ be a sample analogue of $W_t(\mathrm{g}_t)$ for each $t = 1, \ldots, T$ and $W_n(g)$ be a sample analogue of $W_t(\mathrm{g}_t)$: $W_{nt}(\mathrm{g}_t) \equiv n^{-1}\sum_{i=1}^n w_t(\mathrm{g}_t, H_{it})$ and $W_n(g) \equiv \sum_{t=1}^T W_{nt}(\mathrm{g}_t)$. Note that $W_n(g)$ is not a sample analogue of the left hand side of the equation (2.1).

Then, it follows for any $\tilde{g} \in \mathcal{G}$ that

$$\begin{aligned} W(\tilde{g}) - W\left(\hat{g}^S\right) &= W(\tilde{g}) - W_n(\tilde{g}) + W_n(\tilde{g}) - W\left(\hat{g}^S\right) \\ &\leq W(\tilde{g}) - W_n(\tilde{g}) + W_n\left(\hat{g}^S\right) - W\left(\hat{g}^S\right) \\ &\leq 2 \sup_{g \in \mathcal{G}} |W_n(g) - W(g)| \\ &= 2 \sup_{g \in \mathcal{G}} |\{W_{n1}(\mathrm{g}_1) + \cdots + W_{nT}(\mathrm{g}_T)\} - \{W_1(\mathrm{g}_1) + \cdots + W_T(\mathrm{g}_T)\}| \\ &\leq 2 \sum_{t=1}^{T} \sup_{g_t \in \mathcal{G}_t} |W_{nt}(\mathrm{g}_t) - W_t(\mathrm{g}_t)|. \end{aligned} \tag{A.2}$$

The first inequality follows from the fact that $\hat{g}^S$ maximizes $W_n(\cdot)$ over $\mathcal{G}$. Thus, we find that $W^*_{\mathcal{G}} - W\left(\hat{g}^S\right)$ is bounded above from $2\sup_{g \in \mathcal{G}} |W_n(g) - W(g)|$.

For each $t = 1, \ldots, T$, applying Lemma A.1, we have the following result:

$$E_{P \in \mathcal{P}(M,\kappa)}\left[\sup_{g \in \mathcal{G}} |W_{nt}(\mathrm{g}_t) - W_t(\mathrm{g}_t)|\right] \leq C_1 \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}},$$

where C_1 is the same universal constant that appeared in Lemma A.1. Combining this result with (A.2), we have

$$E_{P \in \mathcal{P}(M,\kappa)}\left[\left|W^*_{\mathcal{G}} - W\left(\hat{g}^S\right)\right|\right] \leq 2C_1 \sum_{t=1}^{T} \left\{ \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}} \right\}.$$

(ii) For the Backward DEWM method:
I next provide the proof for Backward DEWM method. For any $\tilde{g} \in \mathcal{G}$, it follows that

$$\begin{aligned} W(\tilde{g}) - W\left(\hat{g}^B\right) =& W(\tilde{g}) - W_n(\tilde{g}) \\ &+ \left\{W_n(\tilde{g}) - W_n\left(\tilde{g}_1, \ldots, \tilde{g}_{T-1}, \hat{g}^B_T\right)\right\} \\ &+ \cdots + \left\{W_n\left(\tilde{g}_1, \hat{g}^B_2, \ldots, , \hat{g}^B_T\right) - W_n\left(\hat{g}^B\right)\right\} \\ &+ W_n\left(\hat{g}^B\right) - W\left(\hat{g}^B\right) \\ \leq& W(\tilde{g}) - W_n(\tilde{g}) + W_n\left(\hat{g}^B\right) - W\left(\hat{g}^B\right) \\ \leq& 2 \sup_{g \in \mathcal{G}} |W_n(g) - W(g)|. \end{aligned}$$

The first inequality follows from the fact that $\hat{g}^B_t$ maximizes $W_n\left(\tilde{g}_1, \ldots, \tilde{g}_{T-1}, \cdot, \hat{g}^B_{t+1}, \ldots, \hat{g}^B_T\right)$ over $\mathcal{G}_t$.

Therefore, following the same argument of the first part of this proof, we have

$$E_{P\in\mathcal{P}(M,\kappa)}\left[\left|W_{\mathcal{G}}^{*}-W\left(\hat{g}^{B}\right)\right|\right]\leq 2C_{1}\sum_{t=1}^{T}\left\{\frac{\gamma_{t}M_{t}}{\prod_{s=1}^{t}\kappa_{s}}\sqrt{\frac{\sum_{s=1}^{t}v_{s}}{n}}\right\},$$

where C_1 is the same universal constant that appeared in Lemma A.1. □

I introduce a definition and lemmas that are used in the proof of Proposition 3.2 and Theorem 4.1. Definition A.1 expresses the complexity of a class of functions. The same definition can be found, for instance, in van der Vaart and Wellner (1996) or Mohri et al. (2012).

Definition A.1. (Rademacher complexity) Let $\mathcal{F}$ be a class of bounded functions mapping from $\mathcal{Z}$ and $S=\{z_1,\ldots,z_n\}$ a fixed sample of size n with elements in $\mathcal{Z}$. Then, the empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample S is defined as:

$$\hat{\mathfrak{R}}_{S}\left(\mathcal{F}\right)=E_{\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_{i}f\left(z_{i}\right)\right],$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. uniform random variables taking values in $\{-1,1\}$ which are called Rademacher variables.

Further, let D denote the distribution according to which samples are drawn. For any integer $n\geq 1$, the Rademacher complexity of $\mathcal{F}$ is the expectation of the empirical Rademacher complexity over all samples of size n drawn according to D:

$$\mathfrak{R}_{S}\left(\mathcal{F}\right)=E_{D^{n}}\left[\hat{\mathfrak{R}}_{S}\left(\mathcal{F}\right)\right].$$

The following lemma relates Rademacher complexity to VC dimension. Its proof can be found in many literatures (e.g., Lugosi (2002); Morhi et al. (2008)).

Lemma A.2. Let $\mathcal{F}$ be a class of bounded functions mapping from $\mathcal{Z}$ such that $\|f\|_{\infty}\leq F$ for all $f\in\mathcal{F}$ and assume its VC-dimension is $v<\infty$. Further, let $S=\{z_1,\ldots,z_n\}$ a fixed sample of size n with elements in $\mathcal{Z}$ and D be the distribution according to which $z_i,\ i=1,\ldots,n$, is drawn. Then, the following holds:

$$\mathfrak{R}_{S}\left(\mathcal{F}\right)\leq F\sqrt{\frac{2v\log\left(\frac{en}{v}\right)}{n}}.$$

The proof of the following lemma can be found, for instance, Morhi et al. (2008).

Lemma A.3. (McDiarmid's Inequality) Let $Z_1, \ldots, Z_n \in \mathcal{Z}^n$ be a set of n independent random variables and g be a mapping from $\mathcal{Z}^n$ to $\mathbb{R}$ such that there exist $c_1, \ldots, c_n > 0$ that satisfy the following conditions:

$$\left| g\left(z_1, \ldots, z_i, \ldots, z_n\right) - g\left(z_1, \ldots, z_i', \ldots, z_n\right) \right| < c_i,$$

for all $i \in \{1, \ldots, n\}$ and any points $\{z_1, \ldots, z_n, z_i'\} \in \mathcal{Z}^{n+1}$. Let $g(S)$ denote $g(Z_1, \ldots, Z_n)$, then the following inequalities hold for all $\epsilon > 0$:

$$\Pr\left[g(S) - E\left[g(S)\right] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right),$$
$$\Pr\left[g(S) - E\left[g(S)\right] \leq -\epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Based on the above lemmas, I provide the proof of Proposition 4.1. The proof follows the similar argument of the proof of Corollary 3.4 of Mohri et al. (2008).

(Proof of Proportion 3.1)
I first prove the first part of the theorem. From the proof of Theorem 3.1, for any $\tilde{g} \in \mathcal{G}$, it follows that

$$W(\tilde{g}) - W\left(\hat{g}^S\right) \leq 2 \sup_{g \in \mathcal{G}} \left| W_n(g) - W(g) \right|. \tag{A.3}$$

We evaluate $|W_n(g) - W(g)|$. Let $S = (Z_1, \ldots, Z_n)$ be a sample and define

$$A(S) \equiv \sup_{g \in \mathcal{G}} \left\{ W(g) - W_S(g) \right\},$$

where $W_S(g)$ is defined as $W_n(g)$ using the sample S. Let me now introduce $S' = (Z_1, \ldots, Z_{n-1}, Z_n')$: a sample that is different from S at the final component.

Then, it follows that

$$\begin{aligned} A(S) - A(S') &= \sup_{g \in \mathcal{G}} \inf_{g' \in \mathcal{G}} \left\{ W(g) - W_S(g) - W(g') + W_{S'}(g') \right\} \\ &\leq \sup_{g \in \mathcal{G}} \left\{ W(g) - W_S(g) - W(g) + W_{S'}(g) \right\} \\ &= \frac{1}{n} \sup_{g \in \mathcal{G}} \left\{ \sum_{t=1}^{T} w_t(\mathrm{g}_t, H_{nt}) - \sum_{t=1}^{T} w_t(\mathrm{g}_t, H'_{nt}) \right\} \\ &\leq \frac{1}{n} \sum_{t=1}^{T} \sup_{g \in \mathcal{G}_t} \left\{ w_t(\mathrm{g}_t, H_{nt}) - \sum_{t=1}^{T} w_t(\mathrm{g}_t, H'_{nt}) \right\} \\ &\leq \frac{1}{n} \sum_{t=1}^{T} \left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right). \end{aligned}$$

The second inequality uses the fact that $\mathcal{G} = \left(\prod_{t=1}^{T} \mathcal{G}_t \right) \cap \tilde{\mathcal{G}} \subset \prod_{t=1}^{T} \mathcal{G}_t$. The last inequality follows from the fact that, under Assumption 2.3, $w_t(\mathrm{g}_t, H_t)$ is bounded from above by $(\gamma_t M_t / 2) / \left(\prod_{s=1}^{t} \kappa_s \right)$.

Since it also follows that $A(S') - A(S) \leq n^{-1} \sum_{t=1}^{T} \left(\gamma_t M_t / \prod_{s=1}^{t} \kappa_s \right)$, applying Lemma A.3 of McDiarmid's inequality, for any $\epsilon > 0$, we get

$$\Pr\left\{ |A(S) - E[A(S)]| \geq \epsilon \right\} \leq \exp\left(\frac{-2n\epsilon^2}{\left\{ \sum_{t=1}^{T} \left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \right\}^2} \right).$$

This is equivalent to the following inequality: for any $\delta \in (0, 1)$,

$$\Pr\left\{ |A(S) - E[A(S)]| \leq \left(\sum_{t=1}^{T} \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \sqrt{\frac{\log(1/\delta)}{2n}} \right\} \geq 1 - \delta. \tag{A.4}$$

Subsequently, we evaluate $E[A]$. Introduce $S' = (Z'_1, \ldots, Z'_n)$ be an independent copy of $S = (Z_1, \ldots, Z_n) \sim P^n$. We denote the probability of S' by $P^{n'}$ and the expectation under $P^{n'}$ by $E_{P^{n'}}(\cdot)$. It follows that

$$\begin{aligned} A(S) &= \sup_{g \in \mathcal{G}} \left\{ E_{P^{n'}}[W_{S'}(g)] - W_S(g) \right\} \\ &\leq E_{P^{n'}} \left[\sup_{g \in \mathcal{G}} \left\{ W_{S'}(g) - W_S(g) \right\} \right]. \end{aligned}$$

Define i.i.d. Rademacher variables $\sigma^n \equiv (\sigma_1, \ldots, \sigma_n)$ such that $\Pr(\sigma_1 = -1) = \Pr(\sigma_1 = 1) = 1/2$ and they are independent of S and S'. Because $\sigma_i \left\{ w(g, Z'_i) - w(g, Z_i) \right\}$ have the same distribution

with $w(g, Z_i') - w(g, Z_i)$, it follows that

$$\begin{aligned}
A(S) &\leq E\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left\{w\left(g, Z_i'\right) - w\left(g, Z_i\right)\right\}\right] \\
&= E\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} \sigma_i \left(w_t\left(\mathrm{g}_t, Z_i'\right) - w\left(\mathrm{g}_t, Z_i\right)\right)\right] \\
&\leq \sum_{t=1}^{T}\left\{E\left[\sup_{\mathrm{g}_t\in\prod_{s=1}^{t}\mathcal{G}_s} \frac{1}{n}\sum_{i=1}^{n} \sigma_i w_t\left(\mathrm{g}_t, Z_i'\right)\right] + E\left[\sup_{\mathrm{g}_t\in\prod_{s=1}^{t}\mathcal{G}_s} \frac{1}{n}\sum_{i=1}^{n} (-\sigma_i)\, w\left(\mathrm{g}_t, Z_i\right)\right]\right\} \\
&= 2\sum_{t=1}^{T} \mathfrak{R}_n\left(w_t\left(\prod_{s=1}^{t}\mathcal{G}_s\right)\right).
\end{aligned}$$

Thus, applying Lemma A.2, we get

$$A(S) \leq \sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\sqrt{\frac{2\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)}{n}}\right]. \tag{A.5}$$

Consequently, combining (A.3), (A.4), and (A.5), for any $\delta \in (0,1)$, it follows with probability at least $1-\delta$ that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}(\kappa,M)}\left[W_{\mathcal{G}}^{*} - W\left(\hat{g}^{S}\right)\right] \\
\leq &\sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\sqrt{\frac{8\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)}{n}}\right] + \left(\sum_{t=1}^{T}\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\sqrt{\frac{2\log(1/\delta)}{n}} \\
= &\sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\left\{\sqrt{8\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)} + \sqrt{2\log(1/\delta)}\right\}\right]/\sqrt{n}.
\end{aligned}$$

For the Backward DEWM method, from the proof of the second part of Theorem 3.1, we have for any $\tilde{g} \in \mathcal{G}$ that

$$W(\tilde{g}) - W\left(\hat{g}^{B}\right) \leq 2\sup_{g\in\mathcal{G}}\left|W_n(g) - W(g)\right|.$$

Therefore, by the same argument of the above proof, we can get the the second result in Proposition 3.1. □

Appendix B.

This section provides the proof of Theorem 4.1. The following lemma will be used in the proof of Theorem 4.1, which is similar to Lemma 2 in Woodworth et al. (2017).

Lemma B.1. Define

$$\mathcal{G}_{\alpha_n}^S \equiv \left\{ g \in \mathcal{G} : \sum_{t=1}^{T} \left(K_{tb} \frac{1}{n} \sum_{i=1}^{n} g_t \left(H_{it} \right) \right) \leq C_b + \alpha_n \text{ for } b = 1, \ldots, B \right\},$$

which is the subset of treatment assignment rules that satisfy the sample budget constraints (3.4). Let g^* be a solution of the constrained maximization problem (3.2). Then, for any $\delta \in (0,1)$, if $\alpha_n > \sqrt{\log\left(B/\delta\right)/\left(2n\right)} \left(\max_{b \in \{1,\ldots,B\}} \sum_{t=1}^{T} K_{tb} \right)$, $g^* \in \mathcal{G}_{\alpha_n}^S$ holds with probability greater than $1-\delta$.

(Proof) It follows that

$$\begin{aligned}
\Pr\left(g^* \notin \mathcal{G}_{\alpha_n}^S\right) &= \Pr\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) - C_b > \alpha_n \text{ for some } b = 1, \ldots, B \right) \\
&\leq \sum_{b=1}^{B} \Pr\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) - C_b > \alpha_n \right) \\
&\leq \sum_{b=1}^{B} \Pr\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) - E\left[\sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) \right] > \alpha_n \right).
\end{aligned}$$

The second inequality follows from the fact that g^* satisfies the population budget/capacity constraints (3.1).
By Hoeffding's inequality, it follows that

$$\Pr\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) - E\left[\sum_{t=1}^{T} K_{tb} g_t^* \left(H_{it} \right) \right] > \alpha_n \right) \leq \exp\left\{ - \frac{2n\alpha_n^2}{\left(\sum_{t=1}^{T} K_{tb} \right)^2} \right\}$$

for each $b = 1, \ldots, B$. Thus, we have

$$\begin{aligned}\Pr\left(g^* \notin \mathcal{G}^S_{\alpha_n}\right) &\leq \sum_{b=1}^{B} \exp\left\{-\frac{2n\alpha_n^2}{\left(\sum_{t=1}^{T} K_{tb}\right)^2}\right\} \\ &\leq B \exp\left\{-\frac{2n\alpha_n^2}{\max_{b\in\{1,\ldots,B\}}\left(\sum_{t=1}^{T} K_{tb}\right)^2}\right\}.\end{aligned}$$

Therefore, if $\alpha_n > \sqrt{\log\left(B/\delta\right)/\left(2n\right)}\left(\max_{b\in\{1,\ldots,B\}}\sum_{t=1}^{T} K_{tb}\right)$, $g^* \in \mathcal{G}^S_{\alpha_n}$ holds with probability greater than $1-\delta$. □

(Proof of Theorem 4.1.)

We use the notation $A \leq_\delta B$ to denote that $A \leq B$ holds with probability at least $1-\delta$.
From the proof of Proposition 3.1, it follows that for any $g \in \mathcal{G}$

$$\begin{aligned}&\sup_{P\in\mathcal{P}(\kappa,M)} \left|W\left(g\right) - W_n\left(g\right)\right| \\ &\leq_\delta \sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\right)\left\{\sqrt{2\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)} + \sqrt{\frac{\log\left(1/\delta\right)}{2}}\right\}\right]/\sqrt{n}, \qquad \text{(B.1)}\end{aligned}$$

and , applying the same argument in proof of Proposition 3.1, we have for each $b = 1, \ldots, B$ that

$$\begin{aligned}&\sup_{P\in\mathcal{P}(\kappa,M)} \left|E\left[\sum_{t=1}^{T} K_{tb}\hat{g}_t^S\left(H_{it}\right)\right] - E_S\left(\sum_{t=1}^{T} K_{tb}\hat{g}_t^S\left(H_{it}\right)\right)\right| \\ &\leq_\delta \sum_{t=1}^{T}\left[K_{tb}\left\{\sqrt{2\left(\sum_{s=1}^{t} v_s\right)\log\left(\frac{en}{\sum_{s=1}^{t} v_s}\right)} + \sqrt{\frac{\log\left(1/\delta\right)}{2}}\right\}\right]/\sqrt{n}. \qquad \text{(B.2)}\end{aligned}$$

By Lemma B.1, if $\alpha_n > \sqrt{\log\left(6B/\delta\right)/\left(2n\right)}\left(\max_{b\in\{1,\ldots,B\}}\sum_{t=1}^{T} K_{tb}\right)$, we have $W_n\left(g^*\right) \leq_{\delta/6} W_n\left(\hat{g}^S\right)$ and $E_n\left[\sum_{t=1}^{T} K_{tb} g_t^*\left(H_{it}\right)\right] \leq_{\delta/6} E_n\left[\sum_{t=1}^{T} K_{tb}\hat{g}_t^S\left(H_{it}\right)\right]$. Combining these results with

(B.1) and (B.2), respectively, it follows that

$$
\begin{aligned}
& W(g^*) \\
\leq_{\delta/6} & W_n(g^*) + \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6/\delta)}{2}} \right\} \right] / \sqrt{n} \\
\leq_{\delta/6} & W_n(\hat{g}_S) + \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6/\delta)}{2}} \right\} \right] / \sqrt{n} \\
\leq_{\delta/6} & W(\hat{g}_S) + 2 \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6/\delta)}{2}} \right\} \right] / \sqrt{n}.
\end{aligned}
$$

and, for each $b = 1, \ldots, B$,

$$
\begin{aligned}
& E_P \left[\sum_{t=1}^{T} K_{tb} g_t^*(H_{it}) \right] \\
\leq_{\delta/(6B)} & E_n \left[\sum_{t=1}^{T} K_{tb} g_t^*(H_{it}) \right] + \sum_{t=1}^{T} \left[K_{tb} \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6B/\delta)}{2}} \right\} \right] / \sqrt{n} \\
\leq_{\delta/(6B)} & E_n \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^S(H_{it}) \right] + \alpha_n + \sum_{t=1}^{T} \left[K_{tb} \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6B/\delta)}{2}} \right\} \right] / \sqrt{n} \\
\leq_{\delta/(6B)} & E_P \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^S(H_{it}) \right] + \alpha_n + 2 \sum_{t=1}^{T} \left[K_{tb} \left\{ \sqrt{2 \left(\sum_{s=1}^{t} v_s \right) \log \left(\frac{en}{\sum_{s=1}^{t} v_s} \right)} + \sqrt{\frac{\log(6B/\delta)}{8}} \right\} \right] / \sqrt{n}.
\end{aligned}
$$

The theorem follows from combining the failure probabilities in the above two equations. □

References

[1] Abbring, J. and Heckman, J. (2007). Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation, in Handbook of Econometrics, Volume 6B, ed. by J. Heckman and E. Leamer, 5145-5303. Elsevier, North-Holland.

[2] Armstrong, T. and Shen, S. (2015). Inference on Optimal Treatment Assignments. Cowles Foundation Discussion Papers 1927RR.

[3] Athey, S. and Wager, S. (2017). Efficient Policy Learning. arXiv preprint arXiv:1702.02896.

[4] Bhattacharya, D. and Dupas, P. (2012). Inferring Welfare Maximizing Treatment Assignment under Budget Constraints. Journal of Econometrics, 167, 168-196.

[5] Chamberlain, G. (2011). Bayesian Aspects of Treatment Choice, in The Oxford Handbook of Bayesian Econometrics, ed. by J. Geweke, G. Koop, and H. vanDijk, 11-39. OxfordUniversity Press, Oxford.

[6] Chakraborty, B. and Murphy, S. (2014). Dynamic Treatment Regimes. Annual Review of Statistics and Its Application, 2014-1, 447-464.

[7] Dehejia, R. (2005). Program Evaluation as a Decision Problem. Journal of Econometrics, 125, 141-173.

[8] Devroye, L., Gyorfi, L., and Lugosi, G. (1996). A Probabilistic Theory of Pattern Recognition. Springer, New-York.

[9] Han, S. (2019). Identification in Nonparametric Models for Dynamic Treatment Effects. Unpublished Manuscript.

[10] Heckman, J., Humphries, J., and Veramendi, G. (2016). Dynamic Treatment Effects. Journal of Econometrics, 191, 276-292

[11] Heckman, J. and Navarro, S. (2007). Dynamic Discrete Choice and Dynamic Treatment Effects. Journal of Econometrics, 136, 341-396.

[12] Hirano, K. and Porter, J. (2009). Asymptotics for Statistical Treatment Rules. Econometrica, 77, 1683-1701.

[13] Kasy, M. (2014). Using Data to Inform Policy. Technical report.

[14] Kitagawa, T. and Tetenov, A. (2018a). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. Econometrica, 86, 591-616.

[15] Kitagawa, T. and Tetenov, A. (2018b). Supplement to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice". Econometrica Supplemental Material, 86.

[16] Kitagawa, T. and Tetenov, A. (2018c). Equality-Minded Treatment Choice. Cemmap Working Paper 71/18.

[17] Kock, A. and Thhyrsgaard, M. (2018). Optimal Sequential Treatment Allocation. arXiv preprint arXiv:1705.09952.

[18] Kolsrud J., Landais, C., Nilsson, P., and Spinnewijn, J. (2018). The Optimal Timing of Unemployment Benefits: Theory and Evidence from Sweden. American Economic Review, 108, 985-1033.

[19] Lechner, M. (2009). Sequential Causal Models for the Evaluation of Labor Market Programs. Journal of Business & Economic Statistics, 27, 71-83.

[20] Lugosi, G. (2002). Pattern Classification and Learning Theory, in Principles of Nonparametric Learning, ed. by L. Györfi, 1–56, Springer, Vienna: .

[21] Lechner, M. and Miquel, R. (2010). Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions. Empirical Economics, 39, 111-137.

[22] Manski, C. (2004). Statistical Treatment Rules for Heterogeneous Populations. Econometrica, 72, 1221-1246.

[23] Mbakop, E. and Tabord-Meehan, M. (2018). Model Selection for Treatment Choice: Penalized Welfare Maximization. arXiv preprint arXiv:1609.03167.

[24] Meyer, B. (1995). Lessons from the U.S. Unemployment Insurance Experiments. Journal of Economic Literature, 33, 91-131.

[25] Moodie E, Chakraborty B, Kramer M. (2012). Q-learning for Estimating Optimal Dynamic Treatment Rules from Observational Data. Canadian Journal of Statistics, 40, 629–645.

[26] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of Machine Learning. The MIT Press, Massachusetts.

[27] Murphy, S. (2003). Optimal Dynamic Treatment Regimes. Journal of the Royal Statistical Society, Series B, 65, 321-366.

[28] Murphy, S. (2005). A generalization Error for Q-learning. Journal of Machine Learning Research. 2005, 6, 1073–1097.

[29] Robins, J. (1989). The Analysis of Randomized and Non-randomized Aids Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies. Health Service Research Methodology: A Focus on AIDS,113-159.

[30] Robins, J., (1997). Causal Inference from Complex Longitudinal Data, in Latent Variable Modeling and Applications to Causality, ed. by. M. Berkane, 69-117, Lecture Notes in Statistics. Springer, New York.

[31] Robins, J. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data.

[32] Rodríguez, J., Saltiel, F., and Urzúa, S. (2018). Dynamic Treatment Effects of Job Training. NBER Working Paper No. 25408.

[33] Stoye, J. (2009). Minimax Regret Treatment Choice with Finite Samples. Journal of Econometrics, 151, 70-81.

[34] Stoye, J. (2012). Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments. Journal of Econometrics, 166, 138-156.

[35] Tetenov, A. (2012). Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria. Journal of Econometrics, 166, 157-165.

[36] Van der Vaart, W. and Wellner, A. (1996). Weak Convergence and Empirical Processes. Springer, New-York.

[37] Vikström, J. (2017). Dynamic Treatment Assignment and Evaluation of Active Labor Market Policies. Labour Economics, 49, 42-54.

[38] Woodworth, B., Gunasekar, S., Ohannessian, M., and Srebro, N. (2017). Learning Non-Discriminatory Predictors. arXiv preprint arXiv:1702.06081.

[39] Zhao, YQ., Zeng, D., Laber, E., and Kosorok, M. (2015). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. Journal of the American Statistical Association, 110, 583-598.