Estimation of Optimal Dynamic Treatment Assignment Rules under Policy Constraints*

Shosei Sakaguchi†

April 8, 2025

Abstract

Many policies involve dynamics in their treatment assignments, where individuals receive sequential interventions over multiple stages. We study estimation of an optimal dynamic treatment regime that guides the optimal treatment assignment for each individual at each stage based on their history. We propose an empirical welfare maximization approach in this dynamic framework, which estimates the optimal dynamic treatment regime using data from an experimental or quasi-experimental study while satisfying exogenous constraints on policies. The paper proposes two estimation methods: one solves the treatment assignment problem sequentially through backward induction, and the other solves the entire problem simultaneously across all stages. We establish finite-sample upper bounds on worst-case average welfare regrets for these methods and show their optimal $n^{-1/2}$ convergence rates. We also modify the simultaneous estimation method to accommodate intertemporal budget/capacity constraints.

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

JEL codes: C22, C44, C54.

*I would like to thank the editor, Stephane Bonhomme, and anonymous board member and referees for their constructive comments and suggestions. I am grateful to Toru Kitagawa, Aleksey Tetenov, Ryo Okui, Jeff Rowley, and participants in various seminars and conferences for their comments and suggestions. This work was supported by JSPS KAKENHI Grant (number 22K20155) and ERC Grant (number 715940).

†Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. Email: sakaguchi@e.u-tokyo.ac.jp.

1 Introduction

Many policies involve dynamics in their treatment assignments. Some policies assign a series of treatments to each individual across multiple stages, such as job training programs consisting of multiple stages (e.g., Lechner, 2009; Rodríguez et al., 2022). Some policies are characterized by when to start/stop consecutive treatment assignment, such as unemployment insurance programs that reduce benefit level after a certain duration (e.g., Meyer, 1995; Kolsrud et al., 2018). Examples of dynamic treatment assignments also include sequential medical interventions, educational programs, and marketing strategies.

When implementing a dynamic policy, policymakers aim to optimize treatment assignments across multiple stages to maximize its social impact. The effects of treatment at each stage are usually heterogeneous with respect to past treatments, intermediate outcomes, and individual characteristics. Hence, to maximize its social impact, treatment assignment to each individual at each stage should depend on the individual's accumulated information up to the corresponding stage.[1]

This paper proposes a statistical decision approach to solve dynamic treatment choice problems using data from an experimental or quasi-experimental study. We assume dynamic unconfoundedness (Robins, 1997), meaning that the treatment assignment at each stage is independent of current and future potential outcomes given the history of treatment assignments and state variables. Under this assumption, we construct an approach to estimate the optimal dynamic treatment regime (DTR)[2] building upon the concept of empirical welfare maximization (EWM) (Kitagawa and Tetenov, 2018b). We call it the dynamic empirical welfare maximization (DEWM). The DEWM approach estimates the optimal DTR by maximizing empirical welfare, a sample mean of propensity-score-weighted outcomes, over a pre-specified class of feasible DTRs. True (estimated) propensity scores are used in the experimental (observational) data setting.

When designing public policy, considering external constraints such as interpretability or fairness in treatment allocation is crucial. The DEWM approach offers favorable

[1]For example, in the context of a sequential job training program, the interest lies in determining which training regimen should be assigned to each individual at each stage, depending on their history of prior training participation, associated labor outcomes, and other observed characteristics. An important question in the context of unemployment insurance policy is when and to whom the insurance level should be reduced, given a recipient's characteristics and past effort toward a job search.

[2]Borrowing from the terminology of statistics literature, we call the dynamic treatment assignment rule DTR.

features to accommodate exogenous policy constraints by restricting the class of feasible DTRs. Moreover, it can be applied to various dynamic treatment choice problems, such as optimal starting and stopping problems, by properly constraining the class of DTRs.

We present two approaches to estimate the optimal DTR. The first estimates the optimal DTR through backward induction, which solves the treatment choice problem from the final to the first stage, supposing at each stage that the optimal treatments are chosen in future stages. The second estimates the optimal DTR simultaneously across all stages, solving the empirical welfare maximization problem at once for the entire DTR.[3]

We reveal that the two approaches complement each other. The backward estimation method is computationally efficient; however, its consistency is ensured only when a pre-specified class of DTRs contains the first-best rule that assigns the best treatment for any history at each stage (except the first stage). Conversely, the simultaneous estimation method consistently estimates the optimal DTR on a pre-specified class of DTRs, irrespective of the feasibility of the first-best rule, at the cost of computational efficiency.

In practical terms, dynamic policies often impose budget or capacity constraints on treatment allocation over time. An ideal DTR should allocate limited resources effectively across stages to maximize welfare. We extend the simultaneous estimation method to problems with intertemporal budget/capacity constraints. We show that the resulting DTR approximately maximizes welfare while satisfying these constraints.

We evaluate the statistical properties of the DEWM approaches in terms of average welfare regret.[4] We derive finite-sample and distribution-free upper bounds on the average welfare regret of the DTR estimated by each of the backward-induction and simultaneous optimization methods. The resulting bounds depend on the sample size n and a measure of complexity of the class of DTRs. Our main theorem shows that the average welfare regret for each method converges to zero at rate $n^{-1/2}$ in the experimental data setting. Furthermore, we show that this convergence rate is optimal.[5] For the budget/capacity constrained problem, we also analyze the excess implementation cost of the estimated DTR relative to the actual budget/capacity. We derive finite-sample and distribution-

[3]Without specifying the direct and indirect effects of the treatment on future outcomes, each approach accounts for these effects within its EWM process.

[4]The average welfare regret is the average welfare loss relative to the maximum welfare achievable in the pre-specified class of DTRs.

[5]To my knowledge, this is the first work to formally show the minimax rate optimality of welfare regrets in estimating optimal DTRs.

free upper bounds on both the welfare regret and the excess cost of the estimated DTR.

Related Literature

This paper contributes to the literature on statistical decision of treatment choice, although much of existing work focuses on the static problem.[6] Policy learning methods by Kitagawa and Tetenov (2018b), Athey and Wager (2021), and Mbakop and Tabord-Meehan (2021) build on the similarity of the empirical welfare maximizing treatment choice and the empirical risk-minimizing classification. Athey and Wager (2021) apply doubly robust estimators to static policy learning, and show that an $n^{-1/2}$-asymptotic upper bound on regret can be achieved even in the observational data setting.

In the dynamic treatment framework, Han (2021) relaxes the sequential randomization assumption, allowing for noncompliance, and studies point identification of the average dynamic treatment effects and optimal non-additive DTR. Han (2023) proposes a method to characterize the sharp partial ordering of the counterfactual welfares of DTRs in an instrumental variable setting. Heckman and Navarro (2007) and Heckman et al. (2016) use exclusion restrictions to identify the dynamic treatment effect, but their focus do not extend to the identification of the optimal DTRs.

Estimation of the optimal DTRs has been widely studied in the biostatistics and statistics literature.[7] Some dominant approaches exist, such as G-estimation (Robins, 1989; Robins et al., 1992) and Q-learning (Murphy, 2005; Moodie et al., 2012). A potential drawback of these approaches is the risk of misspecification of the models relevant to the counterfactual outcomes. By contrast, the DEWM approach does not need to specify any model relevant to the counterfactual outcomes.

Building on the similarity between treatment choice and classification, Zhao et al. (2015) develop estimation methods for the optimal DTRs using the support vector machine with propensity score weighted outcomes. Their approach is computationally efficient because it uses a convex surrogate loss. However, it cannot accommodate exogenous

[6] A partial list of works in that literature includes Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009, 2012), Bhattacharya and Dupas (2012), Chamberlain (2012), Tetenov (2012), Kitagawa and Tetenov (2018b), Athey and Wager (2021), Kitagawa and Tetenov (2018a), Mbakop and Tabord-Meehan (2021), and Kitagawa et al. (2021).

[7] Chakraborty and Moodie (2013), Chakraborty and Murphy (2014), Laber et al. (2014), and Tsiatis et al. (2019) review the developments in this field.

constraints on a class of DTRs.[8] By contrast, our focus lies on estimating the optimal DTRs with exogenous constraints on the class of DTRs, a scenario more commonly encountered in public policy-making.[9] Beyond the results of Zhao et al. (2015), we reveal a tradeoff between imposing constraints on the dynamic treatment choice and the consistency of the backward-induction approach, and formally show the minimax rate optimality of the proposed methods in the context of dynamic treatment choice.

This work is also related to the literature on optimal stopping (e.g., Van Moerbeke, 1976; Rust, 1987; Jacka, 1991; Goel et al., 2017; Nie et al., 2021). Most works in the literature rely either on a known stochastic model (Van Moerbeke, 1976; Rust, 1987; Jacka, 1991) or on a generator of system dynamics (Goel et al., 2017).[10] The methods proposed in our study can estimate the optimal stopping/starting policies from batch data by properly specifying the class of DTRs.

Finally, the dynamic treatment framework we study differs from the bandit problem, for example, studied by Kock and Thyrsgaard (2018). In the bandit problem, different individuals receive treatment at different stages. By contrast, in our dynamic framework, the same individuals progress through different stages and receive sequential treatment interventions across these stages. Additionally, in bandit problems, the treatment effect is explored and exploited across sequential stages, whereas, in our framework, the effects of sequential treatments are estimated before the allocation task.[11][12]

Structure of the Paper

The remainder of this paper proceeds as follows. Section 2 defines the dynamic treatment choice problem. Section 3 presents the two DEWM methods and shows their statistical properties. Section 4 extends the simultaneous estimation method to accommodate

[8] The hinge loss approach in Zhao et al. (2015) loses consistency and computational efficiency, for example, under budget or fairness constraints. Moreover, Laha et al. (2024) show that using a smooth convex surrogate loss or hinge loss in the simultaneous maximization approach can fail to consistently estimate the optimal DTRs.

[9] In the static setting, Kitagawa et al. (2021) show that the surrogate hinge loss approach has consistency in constrained treatment choice problems, which could be extended to our dynamic setting. Their main result for consistency applies when constraints are imposed on the level set of a treatment rule, whereas we consider more general constraints on the functional form of treatment rules.

[10] Nie et al. (2021) propose doubly robust estimation method for the optimal stopping/starting problem.

[11] The bandit problem is an online learning problem, whereas we study an off-line learning problem.

[12] Kallus (2021) study the bandit problem with DTRs, considering the problem of developing and exploiting the optimal DTR in an online setting.

intertemporal budget/capacity constraints. Section 5 proposes estimation methods for the observational data setting. Section 6 shows the results of a simulation study. In Section 7, we apply the proposed methods to the Project STAR (Steps to Achieving Resilience) data, where we estimate an optimal DTR to allocate each student to a class with or without a teacher aide in multiple grades. Section 8 concludes this paper. The Supplemental Appendix (Sakaguchi, 2025) contains some proofs, several extensions, and additional simulation results.

2 Setup

Section 2.1 introduces the dynamic treatment framework, following Robins's dynamic counterfactual outcomes framework (Robins, 1986, 1997). Subsequently, we define the dynamic treatment choice problem in Section 2.2. In this study, we denote by $E_P[\cdot]$ the expectation with respect to a distribution function P.

2.1 Dynamic Treatment Framework

We suppose T ($T < \infty$) stages of binary treatment assignment. Let $D_t \in \{0, 1\}$, for $t = 1, \ldots, T$, denote the binary treatment at stage t. At the end of each stage t, we observe an outcome Y_t. Let X_t be a k-dimensional vector of covariates observed before treatment assignment at stage t. The distribution of X_t may depend on past treatments, outcomes, and covariates. X_1 represents pre-treatment information, containing individuals' demographic characteristics observed before policy implementation. Throughout this paper, for any time-dependent object A_t, we denote by $\underline{A}_t \equiv (A_1, \ldots, A_t)$ a history of the object up to stage t, and denote by $\underline{A}_{s:t} \equiv (A_s, \ldots, A_t)$, for $s \leq t$, a partial history of the object from stage s up to stage t. For example, the treatment history up to stage t is denoted by $\underline{D}_t = (D_1, \ldots, D_t)$. Let $Z \equiv (\underline{D}_T, \underline{Y}_T, \underline{X}_T)$ be the vector containing all observed variables. We define the history in stage t by $H_t \equiv \left(\underline{D}_{t-1}, \underline{Y}_{t-1}, \underline{X}_t\right)$, which is available information for the policymaker when she chooses a treatment assignment at stage t. Note that $H_s \subseteq H_t$ for any $s \leq t$, and $H_1 = (X_1)$. We denote the support of H_t and Z by $\mathcal{H}_t$ and $\mathcal{Z}$, respectively.

We illustrate the dynamic treatment framework with an example of a sequential job

training from Rodríguez et al. (2022). They study the effect of sequential training in Chile's "Franquicia Tributaria" program, where a worker can sequentially participate in multiple training sessions. They consider two stages ($T = 2$) with "$D_1 = 1$" and "$D_2 = 1$" indicating participation in the first and second stages, respectively. Y_1 and Y_2 are the monthly salaries observed after training for each stage. X_1 includes age, gender, initial wage, and education variables, while there are no time-varying covariates X_2.

To formalize our results, we employ the framework of dynamic potential outcomes (Robins, 1986; Murphy, 2003). Let $Y_t(\underline{d}_t)$ denote the potential outcome of $\underline{d}_t \in \{0,1\}^t$ at stage t, representing the outcome for stage t that is realized when the history of treatment up to stage t coincides with $\underline{d}_t$. We implicitly assume that the potential outcomes are not influenced by future treatments, that is, a no-anticipation condition. Given that the covariates X_t may be influenced by past treatments, we define potential covariates as $X_t(\underline{d}_{t-1})$ for each $t \geq 2$ and $\underline{d}_{t-1} \in \{0,1\}^{t-1}$. We denote $X_1(\underline{d}_0) = X_1$ when $t = 1$. The observed outcomes and covariates are defined as $Y_t \equiv Y_t(\underline{D}_t)$ and $X_t \equiv X_t(\underline{D}_{t-1})$, respectively. Denoting $\underline{Y}_t(\underline{d}_t) \equiv (Y_1(d_1), \ldots, \underline{Y}_t(\underline{d}_t))$ and $\underline{X}_t(\underline{d}_{t-1}) \equiv (X_1, X_2(d_1), \ldots, \underline{X}_t(\underline{d}_{t-1}))$, a vector $H_t(\underline{d}_{t-1}) \equiv \left(\underline{d}_{t-1}, \underline{Y}_{t-1}(\underline{d}_{t-1}), \underline{X}_t(\underline{d}_{t-1})\right)$ represents the potential history that is realized when prior treatments are $\underline{d}_{t-1}$. We denote $H_1(\underline{d}_0) = H_1$ when $t = 1$. The observed history is defined as $H_t \equiv H_t(\underline{D}_{t-1})$. Let P be the distribution of all underlying variables $\left(\underline{D}_T, \{\underline{Y}_T(\underline{d}_T)\}_{\underline{d}_T \in \{0,1\}^T}, \{\underline{X}_T(\underline{d}_{T-1})\}_{\underline{d}_{T-1} \in \{0,1\}^{T-1}}\right)$.

From an experimental or observational study, we observe $Z_i \equiv (D_{it}, Y_{it}, X_{it})_{t=1}^T$ for individuals $i = 1, \ldots, n$, where $Y_{it} \equiv Y_{it}(\underline{D}_{it})$ and $X_{it} \equiv X_{it}(\underline{D}_{i,t-1})$ with $Y_{it}(\underline{d}_t)$ and $X_{it}(\underline{d}_{t-1})$ being a potential outcome and covariates for individual i at stage t. We suppose that the vectors of underlying random variables $V_i \equiv \left(\underline{D}_{iT}, \{\underline{Y}_{iT}(\underline{d}_T)\}_{\underline{d}_T \in \{0,1\}^T}, \{\underline{X}_{iT}(\underline{d}_{T-1})\}_{\underline{d}_{T-1} \in \{0,1\}^{T-1}}\right)$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d) with the distribution P. We denote by P^n the joint distribution of $\{V_i : i = 1, \ldots, n\}$.

Let $e_t(d_t, h_t) \equiv \Pr(D_t = d_t \mid H_t = h_t)$ be a propensity score of treatment at stage t given the history up to that point. We suppose that the propensity scores are known in the experimental study but are unknown in the observational study. These settings are considered in Sections 3-4 and Section 5, respectively.

In this study, we suppose that the following assumptions hold.

Assumption 2.1 (Sequential Independence Assumption)**.** *For any $t = 1, \ldots, T$ and*

$\underline{d}_T \in \{0,1\}^T$, $\left(Y_t\left(\underline{d}_t\right), \ldots, Y_T\left(\underline{d}_T\right), X_{t+1}\left(\underline{d}_t\right), \ldots, X_T\left(\underline{d}_{T-1}\right)\right) \perp\!\!\!\perp D_t \mid H_t$ *a.s.*

Assumption 2.2 (Bounded Outcomes)**.** *There exists* $M_t < \infty$ *such that the support of* Y_t *is contained in* $[-M_t/2, M_t/2]$ *for* $t = 1, \ldots, T$.

Assumption 2.1 is known as a dynamic unconfoundedness assumption or sequential/dynamic conditional independence assumption elsewhere, and is commonly used in the literature on dynamic treatment effect analysis (Robins, 1997; Murphy, 2003). This assumption means that the treatment assignment at each stage is independent of the current and future potential outcomes and future covariates conditional on the history up to that point. This is typically satisfied in sequential randomization experiments. In observational studies, this assumption is often controversial but can be satisfied if a sufficient set of confounders is available. Assumption 2.2 is a common assumption in the literature on statistical treatment choice (e.g., Manski, 2004; Stoye, 2009; Kitagawa and Tetenov, 2018b).

2.2 Dynamic Treatment Choice Problem

We aim to develop methods to estimate the optimal DTRs from experimental or observational data with sequential treatment assignment. We denote a treatment rule for each stage t by $g_t : \mathcal{H}_t \mapsto \{0,1\}$, a map from the history up to stage t to a binary treatment. We define the DTR by $g \equiv (g_1, \ldots, g_T)$, a sequence of stage-specific treatment rules. The DTR guides policymakers in selecting treatment for each individual at each stage based on their history up to that point.

We define the counterfactual outcome of a sequence of treatment rules $\underline{g}_t$ for each stage t as $\widetilde{Y}_t\left(\underline{g}_t\right) \equiv \sum_{\underline{d}_t \in \{0,1\}^t} Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\left\{g_s\left(H_s(\underline{d}_{s-1})\right) = d_s\right\}$. This is the counterfactual outcome for stage t that is realized when the sequential treatment assignment up to stage t follows the sequence of treatment rules $\underline{g}_t$.

We then define the welfare of a DTR g by the population mean of a weighted sum of outcomes as follows:

$$W(g) \equiv E_P\left[\sum_{t=1}^{T} \gamma_t \widetilde{Y}_t\left(\underline{g}_t\right)\right] = \sum_{t=1}^{T} E_P\left[\gamma_t \widetilde{Y}_t\left(\underline{g}_t\right)\right], \tag{1}$$

where the weight γ_t, for $t = 1, \ldots, T$, lies in $[0, 1]$ and is chosen by the policy-maker. If the policymaker targets a time-discounted welfare, the weight at each stage is $\gamma_t = \gamma^{T-t}$ with γ being a time-discount factor that lies in $(0, 1)$. If the policymaker targets the outcome for the last stage only, $\gamma_T = 1$ and $\gamma_t = 0$ for all $t \neq T$.

Given the propensity scores $\{e_t(d_t, h_t)\}_{t=1}^T$ and under Assumption 2.1, the welfare function can be identified by the observables only:

$$W(g) = \sum_{t=1}^{T} E_P \left[\frac{\left(\prod_{s=1}^{t} 1\{D_s = g_s(H_s)\}\right) \gamma_t Y_t}{\prod_{s=1}^{t} e_s(D_s, H_s)} \right]. \tag{2}$$

We suppose that the policymaker chooses a DTR from a pre-specified class of feasible DTRs, denoted by $\mathcal{G} \equiv \mathcal{G}_1 \times \cdots \times \mathcal{G}_T$, where $\mathcal{G}_t$ is a class of feasible treatment rules at stage t (i.e., a class of measurable functions $g_t : \mathcal{H}_t \to \{0, 1\}$). Therefore, the ultimate goal of the analysis is to choose an optimal DTR that maximizes the welfare function $W(\cdot)$ over $\mathcal{G}$. [13]

In this study, we constrain the complexity of the class of feasible DTRs in terms of VC-dimension.[14] The following assumption restricts the complexity of the class of feasible DTRs $\mathcal{G}$ in terms of the VC-dimension of $\mathcal{G}_t$ for each $t = 1, \ldots, T$.

Assumption 2.3 (VC-class)**.** *The class of feasible DTRs $\mathcal{G}$ has the form of $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_T$. For $t = 1, \ldots, T$, $\mathcal{G}_t$ is a VC-class of functions and has VC-dimension $v_t < \infty$.*

This assumption restricts the complexity of the class of DTRs $\mathcal{G}$ by restricting the class of feasible treatment rules $\mathcal{G}_t$ for each specific stage. By restricting the complexity, we can select a DTR that is simple to explain/interpret, and can keep estimated DTRs from overfitting the data. Although Assumption 2.3 excludes nonparametric classes of $\mathcal{G}_t$, our framework can accommodate nonparametric approaches by appropriately controlling the growth rate of VC-dimension with the sample size n.

[13] In the context of sequential job training (Lechner, 2009; Rodríguez et al., 2022), $g_t(h_t)$ decides whether an individual with history h_t should receive job training at stage t. The history h_t may include information on past trainings, pre-training and intermediate wages, and educational backgrounds. When Y_t represents the wage at stage t, the optimal DTR is the optimal sequence of treatment rules for determining participation in job training at each stage, to maximize the population mean of the total weighted wages $W(g)$.

[14] The definition of VC-dimension is given in Definition F.1 in the Supplemental Appendix along with some examples.

We can incorporate arbitrary exogenous policy constraints for ethical or political reasons into DTRs by specifying the form of $\mathcal{G}_t$ for each t.[15] Some examples of practically relevant classes of DTRs are linear treatment rules and decision tree rules.

Aside from the constraint on the functional form, we can specify various dynamic treatment choice problems by restricting the intertemporal relationship of treatment rules across stages. Some examples are as follows.

Example 2.1 (Optimal Starting/Stopping Problem)**.** *If the policymaker aims to decide when to start consecutive treatment assignments for each individual, the restriction* $d_s \leq g_t(\cdot)$ *for all* $s \leq t$ *should be imposed on* $\mathcal{G}_t$*. Similarly, the problem of deciding when to stop consecutive treatment assignments can be specified by imposing the restriction* $d_s \geq g_t(\cdot)$ *on* $\mathcal{G}_t$ *for all* $s \leq t$*.*

Example 2.2 (One-Shot Treatment)**.** *If the problem is to decide when to assign a one-shot treatment to each individual, the analyst should impose the restriction* $\sum_{s=1}^{t-1} d_s + g_t(\cdot) \leq 1$ *on* $\mathcal{G}_t$ *for each* t*.*

The VC-dimension of an additionally restricted class does not exceed that of the original class.

Given a class of feasible DTRs $\mathcal{G}$, we assume the following overlap condition holds for the propensity scores $\{e_t(d_t, h_t)\}_{t=1}^T$.

Assumption 2.4 (Overlap Condition)**.** *For* $t = 1, \ldots, T$*, there exists* $\kappa_t \in (0, 1)$ *for which* $\kappa_t \leq e_t(d_t, h_t)$ *holds for any pair* $(d_t, h_t) \in \{0, 1\} \times \mathcal{H}_t$ *such that there exists* $g_t \in \mathcal{G}_t$ *that satisfies* $g_t(h_t) = d_t$*.*

When $\mathcal{G}$ is structurally constrained, Assumption 2.4 is weaker than a common overlap condition that requires the overlap $e_t(d_t, h_t) \in (0, 1)$ for all $(d_t, h_t) \in \{0, 1\} \times \mathcal{H}_t$ and $t = 1, \ldots, T$.[16] This assumption also guides how to design experiments given $\mathcal{G}$; that

[15] Although a treatment rule $g_t(h_t)$ depends on the full-history of covariates $\underline{x}_t$ from stage 1 to t, we can also consider treatment rules that do not depend on the past covariates $\underline{x}_{t-1}$ by restricting the class $\mathcal{G}_t$ such that for any $g_t \in \mathcal{G}_t$, $g_t(h_t) = g_t(h_t')$ for any h_t and h_t' such that $h_t \backslash \underline{x}_{t-1} = h_t' \backslash \underline{x}_{t-1}'$. Similar constraints can also be imposed for the treatment history $\underline{d}_t$ and outcome history $\underline{y}_t$.

[16] For example, in the optimal stopping problem, Assumption 2.4 does not require $e_t(1, h_t) > 0$ for any h_t such that d_s in h_t is equal to 0 for some $s < t$.

is, in an experiment, the treatment d_t does not need to be assigned to individuals with any h_t such that d_t is not achievable by $g_t(h_t)$ for any $g_t \in \mathcal{G}_t$ (i.e., $d_t \neq g_t(h_t)$ for any $g_t \in \mathcal{G}_t$).[17] Assumption 2.4 is satisfied in the experimental data setting, for example, when the treatment D_t is randomly assigned without any dependence on the history H_t.

We denote the highest welfare that is attainable in the class of feasible DTRs $\mathcal{G}$ by

$$W^*_{\mathcal{G}} \equiv \max_{g \in \mathcal{G}} W(g). \tag{3}$$

We consider estimating the optimal DTR that maximizes the welfare $W(\cdot)$ over $\mathcal{G}$ from the sample $\{Z_i : i = 1, \ldots, n\}$. In the subsequent section, we present two methods to estimate the optimal DTR, and show their statistical properties.

3 Dynamic Empirical Welfare Maximization

This section proposes two DEWM methods. One method employs backward induction to solve the dynamic treatment choice problem sequentially from the final to initial stage. The other method involves the simultaneous maximization of $W(\cdot)$ over the entire class of DTRs $\mathcal{G}$ across all stages. The backward-induction approach is computationally efficient; however, we will see that it may not consistently estimate the optimal DTR when $\mathcal{G}_t$ does not contain the first-best treatment rule for all $t \geq 2$. By contrast, the simultaneous maximization method can consistently estimate the optimal DTR irrespective of whether $\mathcal{G}_t$ contains the first-best rule at each stage t, though it is computationally less efficient. We explain the backward-induction and simultaneous-maximization methods in Sections 3.1 and 3.2, respectively.

3.1 Backward Dynamic Empirical Welfare Maximization

We first explain the backward-induction approach. To present the idea, we here suppose that the generative distribution function P is known and the pair $(P, \mathcal{G})$ satisfies Assumptions 2.1 and 2.4.

The backward-induction approach in the population problem proceeds as follows.

[17] For example, in the optimal stopping problem, $d_t = 1$ does not need to be assigned to any individuals who were already untreated (i.e., individuals with $d_s = 0$ for some $s < t$).

First, for the final stage T, we obtain

$$g_T^* \in \arg\max_{g_T \in \mathcal{G}_T} E_P \left[Q_T \left(H_T, g_T(H_T) \right) \right], \tag{4}$$

where $Q_T \left(h_T, d_T \right) \equiv E_P \left[\gamma_T Y_T \mid H_T = h_T, D_T = d_T \right]$ is the conditional mean of the weighted final outcome $\gamma_T Y_T$ given the history h_T and treatment d_T.

Then, recursively, from $t = T - 1$ to 1, we obtain

$$g_t^* \in \arg\max_{g_t \in \mathcal{G}_t} E_P \left[Q_t \left(H_t, g_t(H_t) \right) \right], \tag{5}$$

with $Q_t \left(h_t, d_t \right) \equiv E_P \left[\gamma_t Y_t + Q_{t+1} \left(H_{t+1}, g_{t+1}^*(H_{t+1}) \right) \mid H_t = h_t, D_t = d_t \right]$. The function $Q_t \left(h_t, d_t \right)$ is the action value function for stage t and represents the expected welfare that is realized when the history is h_t, the treatment at stage t is d_t, and the future treatments follow $(g_{t+1}^*, \ldots, g_T^*)$.

Given the propensity scores $\{e_t \left(d_t, h_t \right)\}_{t=1}^T$ and under Assumption 2.1, $E_P \left[Q_t \left(H_t, g_t(H_t) \right) \right]$ can be identified as

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

where

$$q_t \left(Z, g_t; g_{t+1}, \ldots, g_T \right) \equiv \sum_{s=t}^{T} \left\{ \frac{\left(\prod_{\ell=t}^{s} 1 \left\{ D_\ell = g_\ell \left(H_\ell \right) \right\} \right) \gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell \left(D_\ell, H_\ell \right)} \right\}.$$

Hence, the objective function $E_P[Q_t(H_t, g_t)]$ can be expressed in terms of observables only.

Using the inverse propensity score weighting, we propose the estimation method based on the empirical analogue of the above backward induction procedure. We refer to this method as the backward DEWM method. The backward DEWM method first estimates g_T^* by

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

Then, recursively, from $t = T-1$ to 1, the method estimates g_t^* by

$$\hat{g}_t^B \in \arg\max_{g_t \in \mathcal{G}_t} \frac{1}{n} \sum_{i=1}^{n} q_t \left(Z_i, g_t; \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B \right). \tag{6}$$

We denote by $\hat{g}^B \equiv \left(\hat{g}_1^B, \ldots, \hat{g}_T^B\right)$ the DTR obtained from this procedure.

The resulting DTR $\hat{g}^B$ does not necessarily have consistency to the optimal one, $g_{opt}^* \in \arg\max_{g \in \mathcal{G}} W(g)$, unless the class $\mathcal{G}_t$ of treatment rules for each $t \geq 2$ contain the first-best rule that globally maximizes $E_p[Q_t(H_t, g_t(H_t))]$ over all measurable functions of g_t. For any $s < t$, let

$$\widetilde{Y}_t \left(\underline{d}_s, \underline{g}_{(s+1):t} \right) \equiv \sum_{\underline{d}_{(s+1):t} \in \{0,1\}^{t-s}} Y_t(\underline{d}_s, \underline{d}_{(s+1):t}) \cdot \prod_{\ell=s+1}^{t} 1 \left\{ g_\ell \left(H_\ell \left(\underline{d}_{\ell-1} \right) \right) = d_\ell \right\},$$

which is the outcome in stage t that is realized when the treatment assignments from stage 1 to stage s are fixed to $\underline{d}_s$, and the subsequent sequential treatment assignment follows $\underline{g}_{(s+1):t}$.[18] To ensure consistent estimation with a given distribution P, the following assumption requires that the first-best treatment rule is attainable at all but the first stage.

Assumption 3.1 (First-Best Treatment Rule)**.** *For any $t = 2, \ldots, T$, there exists $g_{t,FB}^* \in \mathcal{G}_t$ such that the following holds:*

$$E_P \left[\sum_{s=t}^{T} \gamma_s \widetilde{Y}_s \left(\underline{D}_{t-1}, \underline{g}_{t:s,FB}^* \right) \middle| H_t \right] \geq \max_{d_t \in \{0,1\}} E_P \left[\sum_{s=t}^{T} \gamma_s \widetilde{Y}_s \left(\underline{D}_{t-1}, d_t, \underline{g}_{(t+1):T,FB}^* \right) \middle| H_t \right] \quad a.s.$$

We refer to $g_{t,FB}^*$, which satisfies Assumption 3.1, as the first-best treatment rule at stage t. The first-best rule $g_{t,FB}^*$ always chooses the best treatment for any history h_t given that the first-best rules are followed in the future stages. Assumption 3.1 is satisfied when $\mathcal{G}_t$, $t = 2, \ldots, T$, are rich enough or are correctly specified in the sense that they contain the first-best rule. Note that there is a trade-off between the simplicity of a class of DTRs and the feasibility of Assumption 3.1; while a simpler class of DTRs is often preferable in practice, it is less likely to contain the first-best rule. Assumption 3.1 does not require the class of treatment rules for the first stage $\mathcal{G}_1$ to contain the first-best.

[18] We denote $\widetilde{Y}_t \left(\underline{d}_t, \underline{g}_{(t+1):t} \right) = Y_t \left(\underline{d}_t \right)$ when $s = t$.

When the first-best rule is not attainable in $\mathcal{G}_t$ for some $t \geq 2$, the solution g_s^* of the backward induction for $s \leq t$ does not necessarily correspond to the optimal treatment rule. We illustrate this issue with a simple example in the following remark (and also in the simulation study in Section 6).

Remark 3.1. *Suppose that $T = 2$ and the data-generating process (DGP) P satisfies the following:*

$$\begin{aligned} &E_P[Y_2(1,1)] = 1.0,\ E_P[Y_2(1,0)] = 0.5,\ E_P[Y_2(0,1)] = 0.0,\ E_P[Y_2(0,0)] = 0.6; \\ &D_1 \text{ and } D_2 \text{ are independently distributed as } Ber(1/2). \end{aligned} \tag{7}$$

We set the target welfare to

$$W(g) = E_P\left[\widetilde{Y}_2(g_1, g_2)\right] = E_P\left[\sum_{(d_1,d_2)\in\{0,1\}^2} Y_2(d_1,d_2) \cdot 1\{g_1(H_1) = d_1, g_2(H_2(d_1)) = d_2\}\right].$$

Suppose that the history information are $H_1 = \emptyset$ and $H_2 = (D_1)$.

As an example of a constrained class of DTRs, we consider a class of uniform DTRs; that is $\mathcal{G}_t = \{c_t^0, c_t^1\}$, for $t = 1, 2$, where c_t^0 and c_t^1 denote constant functions such that $c_t^0(h_t) = 0$ and $c_t^1(h_t) = 1$ for any h_t. Under the supposed DGP P, the first-best rule for $t = 2$ is $g_{2,FB}^*(d_1) = d_1$. Hence $\mathcal{G}_2$ does not contain the first-best.

The optimal DTR over the class of constant DTRs is

$$(g_{1,opt}^*, g_{2,opt}^*) = \underset{(g_1,g_2)\in\{c_1^0,c_1^1\}\times\{c_2^0,c_2^1\}}{\arg\max} E\left[\widetilde{Y}_2(g_1, g_2)\right] = (c_1^1, c_2^1),$$

and its welfare is $W(g_{1,opt}^*, g_{2,opt}^*) = E[Y_2(1,1)] = 1.0$. However, the solution (g_1^*, g_2^*) of the backward-induction approach is (c_1^0, c_2^0) because

$$\begin{aligned} \textit{(1st step)} \quad & g_2^* = \underset{g_2\in\{c_2^0,c_2^1\}}{\arg\max} E_P\left[\widetilde{Y}_2(D_1, g_2)\right] = c_2^0; \\ \textit{(2nd step)} \quad & g_1^* = \underset{g_1\in\{c_1^0,c_1^1\}}{\arg\max} E_P\left[\widetilde{Y}_2(g_1, g_2^*)\right] = c_1^0. \end{aligned}$$

Hence, the backward-induction solution $g^* = (c_1^0, c_2^0)$ differs from the optimal solution $g_{opt}^* = (c_1^1, c_2^1)$ over $\mathcal{G}$, resulting in a suboptimal welfare $W(g^*) = E[Y_2(0,0)] = 0.6$.

The above example suggests that when the first-best rule is not feasible in $\mathcal{G}_t$ ($t \geq 2$), the backward-induction solution does not necessarily correspond to the optimal one. This happens because the backward-induction solution g_t^* depends on the DGP P of the observed data, where treatment assignments follow the distribution of the observed treatment assignments (D_1, D_2). This DGP differs from the DGP that arises when the treatment assignments, except for stage t, follow the optimal treatment rules. However, when the first-best rule is feasible in $\mathcal{G}_t$ for each $t \geq 2$, the backward-induction solution g_t^* at each stage corresponds to the first-best rule, under the overlap condition, irrespective of the distribution of (D_1, D_2).

Finally, note that the infeasibility of the first-best rule does not necessarily cause the suboptimality of the backward-induction approach for a fixed DGP. Suppose that the DGP P satisfies the condition (7) with $E_P[Y_2(0,1)] = 0.0$ replaced by $E_P[Y_2(0,1)] = 0.4$. In this case, the backward-induction solution becomes $g^* = (c_1^1, c_2^1)$ and corresponds to the optimal one $g_{opt}^* = (c_1^1, c_2^1)$.[19]

3.2 Simultaneous Dynamic Empirical Welfare Maximization

The second approach is a sample analogue of the entire welfare maximization problem (3). We refer to the proposed method as the simultaneous DEWM method, as it simultaneously estimates the optimal treatment rules across all stages. The method estimates the optimal DTR through the maximization of the sample analogue of (2):

$$\left(\hat{g}_1^S, \ldots, \hat{g}_T^S\right) \in \arg\max_{g \in \mathcal{G}} \sum_{t=1}^{T} \left[\frac{1}{n} \sum_{i=1}^{n} w_t^S(Z_i, \underline{g}_t) \right], \tag{8}$$

where $\underline{g}_t \equiv (g_1, \ldots, g_t)$ is the vector of treatment rules up to stage t and

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

In equation (8), $n^{-1} \sum_{i=1}^{n} w_t^S(Z_i, \underline{g}_t)$ corresponds to the sample analogue of the t-th term in (2). We denote by $\hat{g}^S \equiv \left(\hat{g}_1^S, \ldots, \hat{g}_T^S\right)$ the DTR obtained from this procedure. Theorem

[19] There is also another example. Consider decision rules that rely solely on a discretized version of the history space. In such a scenario, backward induction can still achieve the optimal decision rule within this discretized class, treating the discretized history space as a new set of covariates.

3.6 below shows that this method can consistently estimate the optimal DTR on $\mathcal{G}$ even when $\mathcal{G}_t$ does not contain the first-best rule for some t (i.e., Assumption 3.1 does not hold).

Remark 3.2 (Optimization). *When $\mathcal{G}_t$ $(t = 1, \ldots, T)$ are classes of the linear treatment rules, the optimization problems (6) for the backward DEWM and (8) for the simultaneous DEWM can be formulated as mixed integer linear programming (MILP) problems. Supplemental Appendix H gives details.*

Remark 3.3 (Q-learning). *The Q-learning method is also based on the idea of backward induction (Murphy, 2005; Moodie et al., 2012). In the first step, the method estimates Q-function for stage T, $Q_T^\dagger(h_t, d_t) \equiv E_P[Y_T | H_T = h_T, D_T = d_t]$, through regression of Y_T on (H_T, D_T) and obtain its estimate $\widehat{Q}_T^\dagger(h_t, d_t)$. Then it estimates the optimal treatment rule for stage T as $\hat{g}_T^Q(h_T) = \arg\max_{d_T \in \{0,1\}} \widehat{Q}_T^\dagger(h_t, d_t)$. Recursively, from $t = T-1$ to 1, the method estimates the Q-function (optimal action-value function) for stage t, $Q_t^\dagger(h_t, d_t) \equiv E_P\left[Y_t + \gamma_{t+1} \max_{d_{t+1}} Q_{t+1}^\dagger(h_{t+1}, d_{t+1}) | H_t = h_t, D_t = d_t\right]$, by regressing $Y_t + \gamma_{t+1} \max_{d_{t+1}} \widehat{Q}_{t+1}^\dagger(h_{t+1}, d_{t+1})$ on (H_t, D_t), and obtain its estimate $\widehat{Q}_t^\dagger(h_t, d_t)$.[20] Then it estimates the optimal treatment rule for stage t as $\hat{g}_t^Q(h_t) = \arg\max_{d_t \in \{0,1\}} \widehat{Q}_t^\dagger(h_t, d_t)$. The method yields a DTR $\hat{g}^Q \equiv \left(\hat{g}_1^Q, \ldots, \hat{g}_T^Q\right)$.*

Q-learning is simple to implement and computationally tractable. Moreover, it does not require overlap conditions of propensity scores. However, it requires the correct specification of the Q-functions for consistent estimation of the optimal DTRs, even when experimental data is used. Our proposed methods do not require the specification of the Q-functions; instead, they use the propensity scores. Additionally, while the backward DEWM requires the specified class of DTRs to include the first-best rules, the simultaneous DEWM does not.

Remark 3.4 (Non-Linear Social Welfare). *So far we have considered the linear form (1) of the welfare function. However, some important social welfare criteria (e.g., Gini social welfare (Blackorby and Donaldson, 1978; Weymark, 1981)) are represented by non-linear social welfare functions. In Supplemental Appendix C, we consider the equality-minded*

[20] Linear regression is typically used to estimate the Q-functions.

rank-dependent social welfare functions introduced by Meyer (1995) and Weymark (1981) and studied by Kitagawa and Tetenov (2021):

$$W_\Lambda(F) \equiv \int_0^\infty \Lambda(F(y))dy, \tag{9}$$

where $F(y)$ is the distribution of an outcome and $\Lambda(\cdot) : [0,1] \to [0,1]$ is a non-increasing, non-negative function with $\Lambda(0) = 1$ and $\Lambda(1) = 0$. An important family of social welfare functions represented by (9) is the extended Gini family (Donaldson and Weymark, 1980, 1983; Aaberge et al., 2013): $W_k(F) \equiv \int_0^\infty (1 - F(y))^{k-1}dy$. When $k = 3$, $W_k(F)$ corresponds to the standard Gini social welfare function (Blackorby and Donaldson, 1978; Weymark, 1981): $W_{Gini}(F) = E(Y)(1 - I_{Gini}(F))$ with $I_{Gini}(F) = 1 - (\int_0^1 F^{-1}(\tau) \cdot 2(1 - \tau)d\tau)/E(Y)$.

For any DTR $g = (g_1, \ldots, g_T)$, let $F_g(\cdot)$ denote the distribution of $\sum_{t=1}^T \gamma_t \widetilde{Y}_t(\underline{g}_t)$, and we define the rank-dependent SWF of g by $W_\Lambda(g) \equiv W_\Lambda(F_g)$. Supplemental Appendix C presents a simultaneous DEWM approach to estimate the optimal DTR that maximizes the non-linear social welfare function $W_\Lambda(g)$ over $\mathcal{G}$, and shows its statistical properties.

Remark 3.5 (Multiple Treatment)**.** *We have so far considered DTRs with binary treatment in each stage. Suppose that there are K treatments in each stage. The discussion so far and the presented procedures are easily extendable to the multiple treatment setting by replacing the binary treatment class $\{0, 1\}$ with the multiple one $\{1, \ldots, K\}$. In this case, the treatment rule g_t becomes a map from $\mathcal{H}_t$ to $\{1, \ldots, K\}$. Supplemental Appendix D elaborates on this extension.*

3.3 Statistical Properties

As in much of the literature that follows Manski (2004), we evaluate the statistical properties of the two DEWM methods in terms of the average welfare regret, that is, the average welfare loss relative to the maximum feasible welfare $W^*_{\mathcal{G}}$. Following Kitagawa and Tetenov (2018b), we focus on the non-asymptotic upper bounds of the worst-case average welfare regret, $\sup_{P \in \mathcal{P}(M,\kappa,\mathcal{G})} E_{P^n}\left[W^*_{\mathcal{G}} - W\left(\hat{g}\right)\right]$, where $\mathcal{P}\left(M, \kappa, \mathcal{G}\right)$ is a class of distributions of $\left(\underline{D}_T, \{\underline{Y}_T(\underline{d}_T)\}_{\underline{d}_T \in \{0,1\}^T}, \{\underline{X}_T(\underline{d}_{T-1})\}_{\underline{d}_{T-1} \in \{0,1\}^{T-1}}\right)$ that satisfy Assumptions 2.1, 2.2, and 2.4 with $M \equiv (M_1, \ldots, M_T)'$, $\kappa \equiv (\kappa_1, \ldots, \kappa_T)'$, and a fixed $\mathcal{G}$.

The following theorem provides a finite-sample upper bound on the worst-case average welfare regret and shows its dependence on the sample size n, the VC-dimension of $\mathcal{G}_t$ for each t, and the number of stages T.

Theorem 3.6. *Suppose that Assumptions 2.1, 2.2, and 2.4 hold for any distribution $P \in \mathcal{P}(M, \kappa, \mathcal{G})$ and Assumption 2.3 holds for $\mathcal{G}$.*
(i) For the simultaneous DEWM method, there holds

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

where C is some universal constant.
(ii) Suppose, in addition, that Assumption 3.1 holds for a pair of $\mathcal{G}$ and any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$. Then, for the backward DEWM method, there holds

$$\begin{aligned}\sup_{P \in \mathcal{P}(M,\kappa,\mathcal{G})} E_{P^n}\left[W^*_{\mathcal{G}} - W\left(\hat{g}^B\right)\right] \leq\ & C \sum_{t=1}^{T} \left\{ \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}} \right\} \\ & + C \sum_{t=2}^{T} \frac{2^{t-2}}{\prod_{s=1}^{t-1} \kappa_s} \left(\sum_{s=t}^{T} \left\{ \frac{\gamma_s M_s}{\prod_{\ell=t}^{s} \kappa_\ell} \sqrt{\frac{\sum_{\ell=t}^{s} v_\ell}{n}} \right\} \right),\end{aligned}$$

where C is the same universal constant.

Proof. See Appendix A. □

This theorem shows that the convergence rates of the worst-case average welfare regrets of the two methods are not slower than $n^{-1/2}$. The upper bounds increase with the VC-dimension of $\mathcal{G}_t$, implying that as the candidate treatment rules become more complex, the estimated DTR tends to overfit the data (the distribution of welfare regret becomes more dispersed).[21] The upper bound for the backward DEWM method is greater than that for the simultaneous DEWM method, though neither bound is necessarily sharp. Technically, the difference between these bounds arises from the property of the sequential estimation of the backward DEWM, which leads to additional uncertainty in the estimation.

[21] When the VC-dimension v_t increases with the sample size n, Theorem 3.6 implies that the rate of convergence of the welfare regrets depends on this growth rate.

The next theorem shows a lower bound on the maximum average welfare regret for any data-driven DTR. To present the theorem formally, let $v_{s:t}$, for $s \leq t$, denote the VC-dimension of the following class of indicator functions on $\mathcal{Z}$:

$$\{f(z) = 1\{g_s(h_s) = d_s, \ldots, g_t(h_t) = d_t\} : (g_s, \ldots, g_t) \in \mathcal{G}_s \times \cdots \times \mathcal{G}_t\}.$$

Note that $v_{s:t} \leq \sum_{\ell=s}^{t} v_\ell$ holds (see Lemma A.1).

Theorem 3.7. *Suppose that Assumptions 2.1, 2.2, and 2.4 hold for any distribution $P \in \mathcal{P}(M, \kappa, \mathcal{G})$ and Assumption 2.3 holds for $\mathcal{G}$. Then, for any DTR $\hat{g} \in \mathcal{G}$ as a function of $(Z_1, \ldots, Z_n)$, there holds*

$$\sup_{P \in \mathcal{P}(M,\kappa,\mathcal{G})} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{g})\right] \geq \frac{1}{2}\exp(-4) \max_{t \in \{1,\ldots,T\}} \left\{\gamma_t M_t \sqrt{\frac{v_{1:t}}{n}}\right\}$$

for all $n \geq 16 v_{1:T}$. This result holds irrespective of whether or not Assumption 3.1 additionally holds for a pair of $\mathcal{G}$ and any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$.

Proof. See Supplemental Appendix F. □

This theorem, along with Theorem 3.6, shows that both $\hat{g}^S$ and $\hat{g}^B$ are minimax rate optimal over the class of DGPs $\mathcal{P}(M, \kappa, \mathcal{G})$. Optimality here means that the convergence rates of the upper bounds of the worst-case average welfare regrets in Theorem 3.6 align with the convergence rate of the universal lower bound concerning the sample size n. The convergence rate is also optimal with respect to the VC-dimension v_t for each t. In Theorem 3.7, the maximum of $\gamma_t M_t \sqrt{v_{1:t}/n}$ over $t = 1, \ldots, T$, rather than its summation over $t = 1, \ldots, T$, appears in the lower bound, which is due to the simplicity of the derivation of the lower bound in its proof.

Remark 3.8. *The finite sample optimization problems (6) and (8) are not invariant to adding a constant, which can affect the estimated DTR by manipulating the outcome variables. Following Kitagawa and Tetenov (2018b), we suggest using the demeaned outcomes $Y_{it} - (1/n)\sum_{i=1}^{n} Y_{it}$, instead of the original ones Y_t, in the optimization problems (6) and (8), because it is invariant to adding a constant to the original outcome.*

4 Budget/Capacity Constraints

We consider budget/capacity constraints that limit the proportion of the population receiving treatment. In dynamic treatment policy, these constraints may be imposed intertemporally, meaning that they affect treatment assignment across multiple stages. A policymaker faces an intertemporal budget/capacity constraint when managing a budget that spans across multiple stages or a fixed amount of treatment to distribute over multiple stages.[22] For instance, the job training program studied by Rodríguez et al. (2022) subsidizes training courses at off-site providers across multiple stages, where, when the subsidy budget is limited, the program faces intertemporal budget constraints, limiting the number of individuals participating in training across multiple stages.

Similar to the definition of $\widetilde{Y}_t\left(\underline{g}_t\right)$, we define a counterfactual history as

$$\widetilde{H}_t\left(\underline{g}_{t-1}\right) \equiv \sum_{\underline{d}_{t-1}\in\{0,1\}^{t-1}} H_t(\underline{d}_{t-1}) \cdot \prod_{s=1}^{t-1} 1\left\{g_s\left(H_s(\underline{d}_{s-1})\right) = d_s\right\},$$

which is the counterfactual history in stage t that is realized when the prior treatments $\underline{d}_{t-1}$ are decided by $\underline{g}_{t-1}$. We denote $\widetilde{H}_1\left(\underline{g}_0\right) = H_1$ when $t = 1$. We suppose that the policymaker faces the following B constraints:

$$\sum_{t=1}^{T} K_{tb} E_P\left[g_t\left(\widetilde{H}_t\left(\underline{g}_{t-1}\right)\right)\right] \leq C_b \quad \text{for} \quad b = 1, \ldots, B, \tag{10}$$

where $K_{tb} \in [0, 1]$ and $C_b \geq 0$. As a scale normalization, we assume $\sum_{t=1}^{T} K_{tb} = 1$ for all b. The left-hand side of equation (10) represents the implementation cost of the DTR g, where the weights $K_{1b}, \ldots, K_{Tb}$ represent the relative costs of treatments across stages, and C_b represents the total budget or capacity. If at least two of $K_{1b}, \cdots, K_{Tb}$ take non-zero values, the b-th constraint is an intertemporal budget/capacity constraint; otherwise, the b-th constraint is a temporal one. In the context of the two-stage job training program with an intertemporal budget constraint ($B = 1$), k_{11} and k_{21} represent costs of job training for the first and second stages, respectively, and C_1 represents the

[22] In the static setting, Bhattacharya and Dupas (2012) propose a method to estimate the optimal treatment rule under a budget constraint. As its application, they estimate the optimal allocation policy for subsidies of anti-malaria bed nets under budget constraints.

intertemporal budget of the program.[23] Note that the budget/capacity constraints (10) can be imposed in addition to the constraints considered in the previous sections.

Our aim is to maximize the welfare $W(g)$ under the budget/capacity constraints (10) across the class of feasible DTRs $\mathcal{G}$. The population welfare maximization problem is then formulated as

$$W_{\mathcal{G}}^{*,bdgt} = \max_{g \in \mathcal{G}} W(g) \tag{11}$$
$$\text{s.t.} \sum_{t=1}^{T} K_{tb} E_P \left[g_t \left(\widetilde{H}_t \left(\underline{g}_{t-1} \right) \right) \right] \leq C_b \text{ for } b = 1, \ldots, B.$$

The goal of the analysis is to choose a DTR from $\mathcal{G}$ that maximizes the welfare $W(\cdot)$ subject to the budget/capacity constraints (10).

To this end, we incorporate the sample analogues of the budget/capacity constraints (10) into the simultaneous DEWM.[24] The simultaneous DEWM method with the budget/capacity constraints solves the following problem:[25]

$$\left(\hat{g}_1^{bdgt}, \ldots, \hat{g}_T^{bdgt} \right) \in \arg\max_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} w_t^S \left(Z_i, \underline{g}_t \right) \tag{12}$$
$$\text{s.t.} \sum_{t=1}^{T} K_{tb} \widehat{E} \left[g_t \left(\widetilde{H}_t \left(\underline{g}_{t-1} \right) \right) \right] \leq C_b + \alpha_n \text{ for } b = 1, \ldots, B, \tag{13}$$

where

$$\widehat{E} \left[g_t \left(\widetilde{H}_t \left(\underline{g}_{t-1} \right) \right) \right] \equiv \frac{\sum_{i=1}^{n} \left(\prod_{s=1}^{t-1} 1 \{ D_{is} = g_s (H_{is}) \} \right) g_t (H_{it})}{\sum_{i=1}^{n} \left(\prod_{s=1}^{t-1} 1 \{ D_{is} = g_s (H_{is}) \} \right)}.$$

We denote $\hat{g}^{bdgt} \equiv \left(\hat{g}_1^{bdgt}, \ldots, \hat{g}_T^{bdgt} \right)$.

[23] In reality, the time periods of individuals receiving the treatment would not be aligned. For example, different individuals take job training (for each stage) at different times. In such cases, the formulation (10) of budget constraints can be considered as follows. Suppose that a provider of the treatments (e.g., government) has a fixed budget that can be expended in a fixed fiscal period. The provider (correctly) predicts the number of participants of the program during the fiscal period. We also suppose that the budget can be expended on treatment for any stage for those who participate in the program at any time during the fiscal period. Subsequently, given the budget, the provider can decide the fraction of people who can receive treatment at each stage, as formulated in (10).

[24] We here do not consider the backward DEWM with the budget/capacity constraints because the first-best rule is likely to be unachievable under such constraints.

[25] When $\mathcal{G}_t$ is the class of linear treatment rules for all t, the optimization problem (12) can be formulated as an MILP problem (see Supplemental Appendix H).

The inequality constraints (13) are empirical budget/capacity constraints, where α_n is a tuning parameter dependent on the sample size n. α_n may be either positive or negative, and converges to zero as n increases. As α_n decreases, the empirical budget/capacity constraints become tighter. A sufficiently large value of α_n ensures that the optimal DTR (a solution of (11)) is attainable under the empirical budget/capacity constraint (10) with high probability. In contrast, setting $\alpha_n = 0$ would exclude the optimal DTR from the empirical budget constraint (10) with a non-negligible probability, even with a large sample size.

Subsequently, we evaluate the resulting welfare regret $W_{\mathcal{G}}^{*,bdgt} - W\left(\hat{g}^{bdgt}\right)$ and the budget excess $\sum_{t=1}^{T} K_{tb} E_P\left[\hat{g}_t^{bdgt}\left(H_t\left(\underline{\hat{g}}_{t-1}^{bdgt}\right)\right)\right] - C_b$ of the estimated DTR with high probability, rather than evaluating their expected values, $E_{P^n}\left[W_{\mathcal{G}}^{*,bdgt} - W\left(\hat{g}^{bdgt}\right)\right]$ and $\sum_{t=1}^{T} K_{tb} E_{P^n}\left[E_P\left[\hat{g}_t^{bdgt}\left(H_t\left(\underline{\hat{g}}_{t-1}^{bdgt}\right)\right)\right]\right] - C_b$.[26] We adopt this approach because the actual value of the budget excess is typically of greater concern than its expected value in practice.

The following theorem shows the finite-sample properties of the welfare regret and the budget excess of $\hat{g}^{bdgt}$.

Theorem 4.1. *Suppose that the underlying distribution P satisfies Assumptions 2.1 and 2.2, $\mathcal{G}$ satisfies Assumption 2.3, and that the pair $(P, \mathcal{G})$ satisfies Assumption 2.4. Let $W_{\mathcal{G}}^{*,bdgt}$ be defined in (11) and $\hat{g}^{bdgt}$ be a solution of (12) subject to (13). Let δ be any value in $(0,1)$ and C be the same constant as in Theorem 3.6. Let $k_{(B,n,\delta)} := \sqrt{\log(6B/\delta)/(2n)}$, and $W_{\mathcal{G},\alpha_n}^{*,bdgt}$ be the optimal value of the optimization problem (11) with C_b replaced by $C_b - k_{(B,n,\delta)} + \alpha_n$, assuming that such an optimal value exists. Then the following holds with probability at least $1-\delta$:*

$$\left|W_{\mathcal{G}}^{*,bdgt} - W\left(\hat{g}^{bdgt}\right)\right| \leq \left(W_{\mathcal{G}}^{*,bdgt} - W_{\mathcal{G},\alpha_n}^{*,bdgt}\right) + \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s}\right)\left(2C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{2\log(6/\delta)}\right)\right] \tag{14}$$

[26] Note that $W\left(\hat{g}^{bdgt}\right)$ and $E_P\left[\hat{g}^{bdgt}\left(H_t\left(\underline{\hat{g}}_{t-1}^{S}\right)\right)\right]$ are random variables depending on the random sample $\{Z_i : i = 1, \ldots, n\}$.

and, for any $b \in \{1, \ldots, B\}$,

$$\sum_{t=1}^{T} K_{tb} E_P\left[\hat{g}_t^S\left(\widetilde{H}_t\left(\underline{\hat{g}}_{t-1}^{bgdt}\right)\right)\right] - C_b \leq \frac{1}{\sqrt{n}} \sum_{t=1}^{T}\left[K_{tb}\left(C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log\left(2B/\delta\right)}{2}}\right)\right] + \alpha_n. \tag{15}$$

Proof. See Appendix B. □

Equations (14) and (15) evaluate the welfare regret and budget excess of the estimated DTR over the b-th budget/capacity, respectively. In equation (14), $W_{\mathcal{G}}^{*,bdgt} - W_{\mathcal{G},\alpha_n}^{*,bdgt} \leq 0$ holds when $\alpha_n \leq k_{(B,n,\delta)}$, and $W_{\mathcal{G}}^{*,bdgt} - W_{\mathcal{G},\alpha_n}^{*,bdgt} \geq 0$ holds otherwise. Hence, when we set α_n such that $\alpha_n \leq k_{(B,n,\delta)}$, the results in Theorem 4.1 holds with equation (14) replaced by

$$\left|W_{\mathcal{G}}^{*,bdgt} - W\left(\hat{g}^{bdgt}\right)\right| \leq \frac{1}{\sqrt{n}} \sum_{t=1}^{T}\left[\left(\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s}\right)\left(2C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{2\log\left(6/\delta\right)}\right)\right],$$

where the welfare regret converges to zero as n increases. The theorem suggests that with sufficiently large sample sizes, both the welfare regret and the budget excess are likely to be small, diminishing at the rate of $1/\sqrt{n}$ when α_n is chosen such that $\alpha_n = O(1/\sqrt{n})$ and that $\alpha_n \leq k_{(B,n,\delta)}$.[27]

The tuning parameter α_n decides the strictness of the budget constraint. A smaller α_n implies that the estimated DTR tightly satisfies the budget constraint, as seen in (15), but leads to lower welfare. Conversely, a larger α_n results in less strict adherence to the budget constraint and higher welfare. Thus, the choice of α_n involves a trade-off between maximizing welfare and minimizing budget excess.

We here propose two approaches to choose α_n. When aiming to satisfy the budget constraints with a certain level of budget excesses and a particular probability, we can analytically choose the proper value of α_n through Theorem 4.1. For example, for any $\varepsilon \in (0, 1)$, Theorem 4.1 guarantees that the excess budget for the b-th constraint is equal to or smaller than ε with probability at least $1 - \delta$ when we choose $\alpha_n =$

[27] When $\alpha_n - k_{(B,n,\delta)} \searrow 0$, whether $W_{\mathcal{G}}^{*} - W_{\mathcal{G},\alpha_n}^{*,bdgt}$ in (14) converges to zero depends on the properties of the class of DTR $\mathcal{G}$ and distribution P. When $\mathcal{G}$ consists of a finite number of functions, the maximum welfare subject to budget constraints may not be continuous with respect to the budget under some P, and hence $W_{\mathcal{G}}^{*} - W_{\mathcal{G},\alpha_n}^{*}$ may not converge to zero when $\alpha_n - k_{(B,n,\delta)} \searrow 0$.

$\varepsilon - \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[K_{tb} \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\log\left(2B/\delta\right)/2} \right) \right]$. We can also use cross-validation to choose α_n, wherein the validation data evaluates the welfare and budget excess for each DTR estimated with candidate values of α_n. Note that $k_{(B,n,\delta)}$ is not a tuning parameter to be selected. Theorem 4.1 also guides the selection of the sample size n so that the budget excess is constrained to a certain level with a particular probability.

5 Estimated Propensity Score

We consider the observational data setting where the propensity scores are not known but can be estimated from data. We modify the backward and simultaneous DEWM methods to use the estimated propensity scores, following the e-hybrid EWM rule proposed by Kitagawa and Tetenov (2018b). We also discuss construction of a doubly robust approach for estimating the optimal DTRs with technical exposition and theoretical results presented in Supplemental Appendix E.

Let $\hat{e}_t\left(d_t, h_t\right)$ be an estimated version of the propensity score $e_t\left(d_t, h_t\right)$. For the estimators of the propensity scores, we suppose the following high-level assumption.

Assumption 5.1. *(i) Define*

$$\tau_t\left(\underline{d}_t, H_t\right) \equiv \left\{ \frac{\left(\prod_{s=1}^{t} 1\left\{D_s = d_s\right\}\right) \gamma_t Y_t}{\prod_{s=1}^{t} e_s\left(d_s, H_s\right)} \right\} \text{ and } \hat{\tau}_t\left(\underline{d}_t, H_t\right) \equiv \left\{ \frac{\left(\prod_{s=1}^{t} 1\left\{D_s = d_s\right\}\right) \gamma_t Y_t}{\prod_{s=1}^{t} \hat{e}_s\left(d_s, H_s\right)} \right\},$$

where $\hat{e}_t\left(d_t, H_t\right)$ is an estimated propensity score taking a value in $(0,1)$. For a class of data generating processes $\mathcal{P}_e$, there exists a sequence $\phi_n \to \infty$ such that

$$\sup_{P \in \mathcal{P}_e} \sup_{t \in \{1,\ldots,T\}} \sum_{\underline{d}_t \in \{0,1\}^t} E_{P^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left| \hat{\tau}_t\left(\underline{d}_t, H_{it}\right) - \tau_t\left(\underline{d}_t, H_{it}\right) \right| \right] = O\left(\phi_n^{-1}\right).$$

(ii) Define

$$\eta_t\left(\underline{d}_{t:T}, H_T\right) \equiv \sum_{s=t}^{T} \left\{ \frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = d_\ell\right\}\right) \gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell\left(d_\ell, H_\ell\right)} \right\},$$

$$\hat{\eta}_t\left(\underline{d}_{t:T}, H_T\right) \equiv \sum_{s=t}^{T} \left\{ \frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = d_\ell\right\}\right) \gamma_s Y_s}{\prod_{\ell=t}^{s} \hat{e}_\ell\left(d_\ell, H_\ell\right)} \right\}.$$

For a class of data-generating processes $\widetilde{\mathcal{P}}_e$, there exists a sequence $\xi_n \to \infty$ such that

$$\sup_{P \in \widetilde{\mathcal{P}}_e} \sup_{t \in \{1,\ldots,T\}} \sum_{\underline{d}_{t:T} \in \{0,1\}^{T-t+1}} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\eta}_t\left(\underline{d}_{t:T}, H_{iT}\right) - \eta_t\left(\underline{d}_{t:T}, H_{iT}\right)\right|\right] = O\left(\xi_n^{-1}\right).$$

Note that $E_P\left[\tau_t\left(\underline{d}_t, H_t\right)\right] = E_P\left[\gamma_t Y_t(\underline{d}_t)\right]$ and $E_P\left[\eta_t\left(\underline{d}_{t:T}, H_T\right)\right] = E_P\left[\sum_{s=t}^{T} \gamma_s Y_s(\underline{d}_s)\right]$ hold under Assumption 2.1 and $n^{-1}\sum_{i=1}^{n} \hat{\tau}_t\left(\underline{d}_t\right)$ and $n^{-1}\sum_{i=1}^{n} \hat{\eta}_t\left(\underline{d}_{t:T}\right)$ are estimators of these, respectively. We do not explore lower-level conditions that satisfy Assumption 5.1. When the propensity scores are consistently estimated with parametric estimators, they are estimated at rate $n^{-1/2}$.

When the estimated propensity scores are used, the backward DEWM method solves the following problem, recursively, from $t = T$ to 1:

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

$$\text{with} \quad \hat{q}_t\left(h_t, g_t; g_{t+1}, \ldots, g_T\right) \equiv \sum_{s=t}^{T} \left\{\frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = g_\ell\left(H_\ell\right)\right\}\right) \cdot \gamma_s Y_s}{\prod_{\ell=t}^{s} \hat{e}_\ell\left(D_\ell, H_\ell\right)}\right\}.$$

We denote by $\hat{g}_e^B \equiv \left(\hat{g}_{1,e}^{B}, \ldots, \hat{g}_{T,e}^{B}\right)$ the DTR obtained by this procedure.

Similarly, the simultaneous DEWM method solves the following problem:

$$\left(\hat{g}_{1,e}^{S}, \ldots, \hat{g}_{T,e}^{S}\right) \in \arg\max_{g \in \mathcal{G}} \sum_{t=1}^{T}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{w}_t^S(Z_i, \underline{g}_t)\right]$$

where $\hat{w}_t^S(Z_i, \underline{g}_t) \equiv \left\{\left(\prod_{s=1}^{t} 1\left\{D_{is} = g_s\left(H_{is}\right)\right\}\right) \cdot \gamma_t Y_{it}\right\} / \left\{\prod_{s=1}^{t} \hat{e}_s\left(D_{is}, H_{is}\right)\right\}$ uses the estimated propensity scores. We denote the resulting DTR by $\hat{g}_e^S \equiv \left(\hat{g}_{1,e}^{S}, \ldots, \hat{g}_{T,e}^{S}\right)$.

The following theorem shows the uniform convergence rate bounds on the worst-case average welfare regret for the two estimation methods.

Theorem 5.1. *Suppose that Assumptions 2.1, 2.2, and 2.4 hold for any distribution $P \in \mathcal{P}\left(M, \kappa, \mathcal{G}\right)$ and Assumption 2.3 holds for $\mathcal{G}$.*

(i) Suppose further that Assumption 5.1 (i) holds for any distribution $P \in \mathcal{P}_e$. For the

Simultaneous DEWM method, there holds

$$\sup_{P\in\mathcal{P}_e\bigcap\mathcal{P}(M,\kappa,\mathcal{G})} E_{P^n}\left[W^*_{\mathcal{G}} - W\left(\hat{g}^S_e\right)\right] \leq C\sum_{t=1}^{T}\left\{\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\sqrt{\frac{\sum_{s=1}^{t} v_s}{n}}\right\} + O\left(\phi_n^{-1}\right),$$

where C is the same universal constant as that introduced in Theorem 3.6.
(ii) Suppose that Assumption 5.1 (ii) holds for any distribution $P \in \widetilde{\mathcal{P}}_e$ and Assumption 3.1 holds for a pair $(P, \mathcal{G})$ for any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$. Then, for the backward DEWM method, there holds

$$\begin{aligned}\sup_{P\in\widetilde{\mathcal{P}}_e\bigcap\mathcal{P}(M,\kappa,\mathcal{G})} E_{P^n}\left[W^*_{\mathcal{G}} - W\left(\hat{g}^B_e\right)\right] \leq\ & C\sum_{t=1}^{T}\left\{\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\sqrt{\frac{\sum_{s=1}^{t} v_s}{n}}\right\} \\ & + C\sum_{t=2}^{T}\frac{2^{t-2}}{\prod_{s=1}^{t-1}\kappa_s}\left(\sum_{s=t}^{T}\left\{\frac{\gamma_s M_s}{\prod_{\ell=t}^{s}\kappa_\ell}\sqrt{\frac{\sum_{\ell=t}^{s} v_\ell}{n}}\right\}\right) \\ & + O\left(\xi_n^{-1}\right).\end{aligned}$$

Proof. See Supplemental Appendix F. □

The theorem implies that the convergence rate of the worst-case average regret for each method depends on that of the propensity scores' estimators. If the propensity scores are correctly specified and parametrically estimated, both methods achieve the optimal $n^{-1/2}$-convergence rate of the worst-case average regret.

When the propensity scores are not consistently estimated, the IPW approaches do not consistently estimate the optimal DTRs. We hence propose a doubly robust approach; it is robust to misspecification of either propensity scores or models relevant to outcomes, and can achieve the optimal $n^{-1/2}$-rate of the welfare regret even when nuisance components are nonparametrically estimated. The following remark briefly discusses this approach while technical exposition and theoretical results are presented in Supplemental Appendix E.

Remark 5.2 (Doubly Robust Approach)**.** *For the static treatment choice problem, Athey and Wager (2021) and Zhou et al. (2023) show that using the augmented inverse probability weighting (AIPW) estimator of the welfare function can improve the convergence rate*

of the welfare regret relative to the e-hybrid EWM rule.[28] *In the dynamic setting, we consider extension of the simultaneous maximization approach to doubly robust approach.*[29] *The approach presented in Supplemental Appendix E combines the estimated propensity scores and estimators of Q-functions,*

$$Q_t^{\underline{g}_{(t+1):T}}(h_t, d_t) \equiv E_P\left[\gamma_t Y_t + \sum_{s=t+1}^{T} \gamma_s \widetilde{Y}_s(\underline{D}_t, \underline{g}_{(t+1):s}) \middle| H_t = h_t, A_t = d_t\right],$$

to construct an AIPW estimator of the welfare function $W(g)$, and then maximizes it over $\mathcal{G}$ to estimate the optimal DTR. The cross-fitting is also used. This approach consistently estimates the optimal DTR if either a propensity score or Q-function for each stage is consistently estimated. The results in Supplemental Appendix E show that the welfare regret $W_{\mathcal{G}}^* - W(\hat{g}^{AIPW})$ converges to 0 with the optimal rate of $n^{-1/2}$ under mild conditions on the convergence rates of the estimators of the propensity scores and Q-functions.

This approach, however, faces an optimization challenge. Since $Q_t^{\underline{g}_{(t+1):T}}(h_t, d_t)$ is specific to a sequence of treatment rules $\underline{g}_{(t+1):T}$, when we apply this approach, we have to estimate $\left\{Q_t^{\underline{g}_{(t+1):T}}(h_t, d_t)\right\}_{t=1,\ldots,T}$ for every possible DTR g in $\mathcal{G}$. This is computationally challenging unless the class of DTRs is sufficiently small (e.g., a finite class of a moderate number of DTRs).[30]

6 Simulation Study

We conduct a simulation study to examine the finite sample performance of the proposed methods. We compare the performance of backward DEWM, simultaneous DEWM, and Q-learning.

We consider DGPs that consist of two stages of treatment assignment (D_1, D_2), associated potential outcomes $\left(Y_1(d_1), Y_2(d_1, d_2)\right)_{\{d_1,d_2\}\in\{0,1\}^2}$, and a covariate X_1 observed

[28] Nie et al. (2021) extend this approach to the problem of optimal starting/stopping decision.

[29] Sakaguchi (2024) propose a doubly robust method with backward induction for estimating optimal DTRs.

[30] When covariates are exongenous and intermediate outcomes are not used, a doubly robust approach with lower computational cost can be constructed. Supplemental Appendix E.2 gives details.

at the first stage. The potential outcomes are generated as

$$Y_1(d_1) = \phi_{01} + \phi_{11} X_1 + (\psi_{01} + \psi_{11} X_1) d_1 + U_1,$$
$$Y_2(d_1, d_2) = \phi_{02} + \phi_{12} Y_1(d_1) + \left(\psi_{02} + \psi_{12} d_1 + \sum_{j=1}^{3} \psi_{j+1,2} (Y_1(d_1))^j \right) d_2 + U_2$$

for $(d_1, d_2) \in \{0,1\}^2$. We consider three DGPs labeled DGPs 1-3. In all the DGPs, X_1, U_1, and U_2 are independently drawn from $N(0,1)$; D_1 and D_2 are independently drawn from $Ber(1/2)$; and $(\phi_{01}, \phi_{11}, \psi_{01}, \psi_{11}) = (0.5, -1.0, 1.0, 1.5)$ and $(\phi_{02}, \phi_{12}, \psi_{02}, \psi_{12}) = (0.5, 0.5, 0.5, 0.5)$. Regarding the other parameters, we set $(\psi_{22}, \psi_{32}, \psi_{42}) = (0,0,0)$ in DGP1, $(\psi_{22}, \psi_{32}, \psi_{42}) = (1,0,0)$ in DGP2, and $(\psi_{22}, \psi_{32}, \psi_{42}) = (0.3, 0.3, -0.4)$ in DGP3. We set the target welfare to maximize as $W(g_1, g_2) = E_P[Y_2(g_1, g_2)]$. The treatment effect of D_2 does not depend on the past outcome in DGP1, but it does in DGPs 2 and 3.

For the backward and simultaneous DEWM methods, we use a class of DTRs $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$ that consists of the following classes of linear treatment rules:

$$\mathcal{G}_1 = \left\{ 1\left\{ (1, X_1)' \boldsymbol{\beta}_1 \geq 0 \right\} : \boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11})' \in \mathbb{R}^2 \right\},$$
$$\mathcal{G}_2 = \left\{ 1\left\{ (1, D_1, Y_1)' \boldsymbol{\beta}_2 \geq 0 \right\} : \boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22})' \in \mathbb{R}^3 \right\}.$$

$\mathcal{G}_2$ contains the first-best rule under DGPs 1 and 2 but not under DGP3. Thus, the backward DEWM method can consistently estimate the optimal DTR under DGPs 1 and 2 but cannot under DGP3. We solve the optimization problems for each DEWM method through MILPs as discussed in Remark 3.2.

For Q-learning, we assume that the conditional outcomes are specified as

$$E[Y_1 \mid H_1, D_1; \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1] = \alpha_{01} + \alpha_{11} X_1 + (\gamma_{01} + \gamma_{11} X_1) D_1,$$
$$E[Y_2 \mid H_2, D_2; \boldsymbol{\alpha}_2, \boldsymbol{\gamma}_2] = \alpha_{02} + \alpha_{12} Y_1 + (\gamma_{02} + \gamma_{12} D_1 + \gamma_{22} Y_1) D_2,$$

where $\boldsymbol{\alpha}_t' = (\alpha_{0t}, \alpha_{1t})'$ for each $t = 1, 2$, $\boldsymbol{\gamma}_1' = (\gamma_{01}, \gamma_{11})'$, and $\boldsymbol{\gamma}_2' = (\gamma_{02}, \gamma_{12}, \gamma_{22})'$. This specification is correct under DGPs 1 and 2 but is not under DGP3.

Table 1 presents the results of 500 simulations with sample sizes $n = 200$, 500, and 800. The table shows the mean and median welfare achieved by each estimated DTR calculated with 3,000 observations randomly drawn from the same DGP. The results show

that Q-learning performs better than the backward and simultaneous DEWM methods in DGPs 1 and 2 in terms of the population mean welfare. However, both the backward and simultaneous DEWM methods exhibit superior performance to Q-learning in DGP3, where the outcome model used by Q-learning is misspecified. In DGPs 1 and 2, the backward and simultaneous DEWM methods demonstrate similar welfare performance. However, in DGP3, the simultaneous DEWM method achieves higher welfare than the backward DEWM method. Table 1 also presents the average CPU time to calculate DTR per simulation iteration.[31] The simultaneous DEWM takes the longest time but remains feasible at these scales of simulation. Supplemental Appendix G provides additional simulation results for DGPs where D_1 and D_2 are not independent of U_1 and U_2 (i.e., the sequential independence assumption (Assumption 2.1) does not hold).

Table 1: Monte Carlo simulation results

		n=200				n=500				n=800			
	DGP	Mean	Med	SD	Time	Mean	Med	SD	Time	Mean	Med	SD	Time
Q-learning	1	2.27	2.27	0.04	0.02	2.28	2.28	0.04	0.01	2.28	2.28	0.05	0.01
B-DEWM	1	2.05	2.13	0.24	0.93	2.18	2.22	0.14	5.52	2.21	2.24	0.15	14.49
S-DEWM	1	1.99	2.11	0.29	6.25	2.15	2.20	0.17	41.35	2.21	2.23	0.11	97.34
Q-learning	2	3.97	3.97	0.07	0.01	3.98	3.98	0.07	0.01	3.98	3.98	0.07	0.01
B-DEWM	2	3.52	3.73	0.54	0.75	3.72	3.79	0.37	4.84	3.80	3.81	0.22	12.07
S-DEWM	2	3.50	3.68	0.53	4.52	3.72	3.77	0.33	29.10	3.78	3.81	0.23	64.73
Q-learning	3	1.64	1.63	0.11	0.01	1.62	1.62	0.09	0.01	1.62	1.61	0.09	0.02
B-DEWM	3	1.77	1.80	0.12	0.75	1.74	1.76	0.11	3.85	1.63	1.62	0.11	13.20
S-DEWM	3	1.86	1.89	0.15	5.00	1.85	1.87	0.15	34.71	1.73	1.72	0.15	92.81

Notes: Mean and Med represent the mean and median of the population mean welfares achieved by the estimated DTRs across the simulations, respectively. SD is the standard deviation of the population mean welfares across the simulations. The population mean welfare is calculated using 3,000 observations randomly drawn from the corresponding DGP. B-DEWM and S-DEWM represent the Backward and Simultaneous DEWM methods, respectively. The columns of "Time" show average CPU time to estimate DTR per iteration for each method, DGP, and sample size.

7 Empirical Application

We apply the proposed methods to data from Project STAR (Achilles et al., 2008). Specifically, we use the replication dataset from Krueger (1999), which was used in Angrist and Pischke (2009) and obtained from the MHE Data Archive (Angrist, 2009). In this experimental project, of 1,346 kindergarten students not belonging to small classes, 672 students were randomly assigned to regular-size classes with a full-time teacher aide, and the others

[31] We use Julia version 1.11.1 with Gurobi Optimizer version 12.0.1. The hardware is 13th Gen Intel(R) Core(TM) i9-13900 2.00 GHz.

were allocated to regular-size classes without a teacher aide. Upon their progression to grade 1, the enrolled students were randomly shuffled into regular-class size classes with or without a teacher aide and remained in the allocated classes until the end of grade 3.

We study optimal allocation of students to two types of classes, regular-size classes with or without a teacher aide, in grades K and 1, based on their socioeconomic information and intermediate academic performance.[32] We aim to maximize the population average of scores from mathematics test that students took at the end of grade 1.[33] We set the first and second stages ($t = 1$ and 2) to grades K and 1, respectively. The treatment variable D_t, for $t = 1, 2$, takes the value one if the student is assigned to a class with a teacher aide at stage t and zero otherwise. The potential intermediate outcome $Y_1(d_1)$ and final outcome $Y_2(d_1, d_1)$ represent the mathematics test scores at the end of grades K and 1 for treatments d_1 and d_2.

Since employing a full-time teacher aide incurs costs, we incorporate this expense into our class allocation problem. As our outcome measure is test scores, we convert the monetary cost of a full-time teacher aide per student into equivalent test score units, following the discussion in Krueger (1999, Section IV). Supplemental Appendix I provides a detailed discussion, where considering both the monetary cost of a full-time teacher aide and the present value (at age six) of lifetime earnings due to test score increments, we estimate that the costs per student in grades K and 1 are equivalent to 0.782 and 0.806 points, respectively, on the grade 1 mathematics test.

Letting $c_1 = 0.782$ and $c_2 = 0.806$ represent the costs of the teacher aide per student measured by the test score, we consider $Y_2^c(d_1, d_2) := Y_2(d_1, d_2) - c_1 d_1 - c_2 d_2$ to represent the individual welfare contribution given the cost of teacher aide. We aim to maximize welfare defined as

$$W(g_1, g_2) = E\left[\sum_{(d_1,d_2)\in\{0,1\}^2} Y_2^c(d_1, d_2) \cdot 1\left\{g_1(H_1) = d_1, g_2(H_2(d_1)) = d_2\right\}\right],$$

[32] We focus on allocating students to regular-size classes with a teacher aide, rather than small-size classes, because the allocation of students to regular-size classes with or without a teacher aide in the experiment matches the sequential randomization design; however, the allocation to small-size classes in the experiment does not.

[33] We focus on the test score at the end of grade 1 rather than at the end of grade 3 because we found attending a class with a teacher aide in kindergarten has little effect on the test score at the end of grade 3 even when treatment effect heterogeneity is considered.

which represents the average of the total test score at the end of grade 1 with subtraction of the cost.

The socioeconomic information we use for treatment choice are the qualification for free or reduced-price school lunches and school location (rural or non-rural).[34] A binary variable $X_{Lunch,t}$ takes 1 if the student is eligible for free or reduced-price school lunch at stage t and 0 otherwise. A binary variable $X_{Rural,t}$ takes 1 if the student attends a school located in a rural area at stage t and 0 otherwise. The outcomes Y_1 and Y_2 are demeaned in accordance with Remark 3.8.

We employ a set of class-allocation policies represented by treatment rules $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$, where $\mathcal{G}_1$ and $\mathcal{G}_2$ constitute a class of linear treatment rules:[35]

$$\mathcal{G}_1 = \left\{ 1\left\{\beta_0 + \beta_1 x_{Lunch,1} + \beta_2 x_{Rural,1} \geq 0\right\} : \beta_1 \geq 0,\ (\beta_0, \beta_2)' \in \mathbb{R}^2 \right\},$$
$$\mathcal{G}_2 = \left\{ 1\left\{\gamma_0 + \gamma_1 x_{Lunch,2} + \gamma_2 x_{Rural,2} + \gamma_3 (1 - d_1) y_1 + \gamma_4 d_1 y_1 \geq 0\right\} : \gamma_1 \geq 0,\ (\gamma_0, \gamma_2, \gamma_3, \gamma_4)' \in \mathbb{R}^4 \right\}.$$

In the formulations of $\mathcal{G}_1$ and $\mathcal{G}_2$, the coefficients of $x_{Lunch,1}$ and $x_{Lunch,2}$ are constrained to be non-negative. This ensures that students eligible for free or reduced-price school lunches are not less likely to be allocated to a class with a teacher aide, given that the other information is fixed. The interaction terms $(1 - d_1) y_1$ and $d_1 y_1$ in $\mathcal{G}_2$ enable the eligibility score to evaluate the intermediate outcome differently based on class allocation at kindergarten. We solve the optimization problems for the backward and simultaneous DEWM through MILPs as discussed in Remark 3.2.

For a DTR $g \in \mathcal{G}$, we define the welfare gain of g as $W(g) - E[Y_2^c(0,0)]$, the welfare increase achieved by allocating students according to the DTR g rather than allocating every student to regular classes without teacher aides at all stages.[36] Applying the backward and simultaneous DEWM methods, we estimate the optimal DTR over $\mathcal{G}$ and its welfare gain as well as the treatment ratio at each stage. To avoid overfitting biases, we

[34] While we have access to student information such as sex and race, using such information in treatment choice is discriminatory and prohibited.

[35] It is testable whether $\mathcal{G}_2$ contains the first-best rule or not. For example, we can estimate the first-best rule for the second stage as $\hat{g}_2^{*,FB}(h_2) = 1\{\hat{\tau}_2(h_2) \geq 2\}$ with $\hat{\tau}_2(h_2)$ being a (nonparametric) estimator of the conditional average treatment effect $\tau_2(h_2) = E[Y_2(D_1, 1) - Y_2(D_1, 0) | H_2 = h_2]$. We can then check whether $\mathcal{G}_2$ contains the first-best rule by examining if the optimal policy in $\mathcal{G}_2$ achieves the same expected outcome value as $\hat{g}_2^{*,FB}(h_2)$.

[36] Two factors enhance the students' academic achievement in classroom allocation: optimal matching between each student and classroom type (with or without a teacher aide) and peer effects among students. The optimal DTR considered here exploits the former but not the latter, as it does not utilize peer effects among students to determine classroom allocation.

adopt two-fold random sample splitting with a fixed seed: one third of the sample is used as the training set to estimate the optimal DTRs, and the remaining is used as the test set to estimate the welfare gains and treatment ratios.

The DTR estimated by the backward DEWM method is $\hat{g}^B = \left(\hat{g}_1^B, \hat{g}_2^B\right)$ where $\hat{g}_1^B(h_1) = 1$ and $\hat{g}_2^B(h_2) = 1\left\{-0.059 + 0.528x_{Rural,2} - 0.105\left(1 - d_1\right)y_1 + 0.307d_1y_1 \geq 0\right\}$. The DTR estimated by the simultaneous DEWM method is $\hat{g}^S = \left(\hat{g}_1^S, \hat{g}_2^S\right)$ where $\hat{g}_1^S(h_1) = 1\left\{x_{Rural,1} = 1\right\}$ and $\hat{g}_2^S(h_2) = 1\left\{1.0 + 0.153d_1y_1 \geq 0\right\}$. $\hat{g}_1^B$ assigns every student to a class with a teacher aide in grade K, while $\hat{g}_1^S$ assigns only students in rural areas to classes with teacher aides in grade K. Under both treatment rules $\hat{g}_2^B$ and $\hat{g}_2^S$ for grade 1, a student who attends a class with a teacher aide and attains a high test score in grade K is more likely to be assigned to a class with a teacher aide in grade 1.

Table 2 reports the estimated welfare gains and shares of the population to be treated at each stage for the estimated DTRs $\hat{g}^B$ and $\hat{g}^S$ and three uniform DTRs $(g_1, g_2) = (1,0), (0,1), (1,1)$.[37] For example, the DTR $(g_1, g_2) = (1,0)$ assigns every student to a class with a teacher aide in kindergarten but assigns none in grade 1. The results indicate that both the backward and simultaneous DEWM methods lead to higher welfare gains than all the uniform DTRs. The welfare gains are not very different between the two DEWM methods, possibly because the policy class $\mathcal{G}_2$ for the second stage includes the first-best rule (i.e., Assumption 3.1 holds).

Table 2: Estimated welfare gains

	Share of population to be treated		
Dynamic treatment regime	1st stage	2nd stage	Estimated welfare gain
$(g_1, g_2) = (1,0)$	1	0	2.59
$(g_1, g_2) = (0,1)$	0	1	3.94
$(g_1, g_2) = (1,1)$	1	1	0.80
$\left(\hat{g}_1^B, \hat{g}_2^B\right)$	1.0	0.51	8.86
$\left(\hat{g}_1^S, \hat{g}_2^S\right)$	0.63	0.76	8.80

Notes: The standard deviation of $Y_2(0,0)$ in the sample is 39.65. We use the two-fold sample splitting with a fixed seed. The training sample is used to estimate the DTRs $\hat{g}^B$ and $\hat{g}^S$. The test sample is used to estimate shares of population to be treated and welfare gains of $\hat{g}^B$ and $\hat{g}^S$.

Next, we consider the decision of when each student should begin attending a class with a teacher aide. The motivation for this exercise comes from two main factors. First, many educational policies involve optimal timing decisions (e.g., when to begin foreign language

[37] With some abuse of notation, we denote by $(g_1, g_2) = (d_1, d_2)$ the uniform DTR that allocates every student to class types d_1 and d_2 in stages 1 and 2, respectively.

instruction). Second, the DTR for the optimal starting problem is simpler to implement than for a more general problem, as the administrator does not need to transfer any students from a class with a teacher aide to a class without one in subsequent grades. Moreover, this approach can help avoid situations where students with a teacher aide in one grade lose that support in the following grade. Under the constraint of the optimal starting problem, the DTR estimated by the backward DEWM method is $\hat{g}^B = \left(\hat{g}_1^B, \hat{g}_2^B\right)$ with $\hat{g}_1^B(h_1) = 0$ and $\hat{g}_2^B(h_2) = 1\left\{-0.059 + 0.528x_{Rural,2} - 0.105\left(1 - d_1\right)y_1 + 0.307d_1y_1 \geq 0\right\}$; the DTR estimated by the simultaneous DEWM method is $\hat{g}^S = \left(\hat{g}_1^S, \hat{g}_2^S\right)$ with $\hat{g}_1^S(h_1) = 1\left\{x_{Rural,1} = 1\right\}$ and $\hat{g}_2^S(h_2) = 1$.

Table 3 reports the estimated welfare gains and shares of population to be treated by $\hat{g}^B$ and $\hat{g}^S$ and two uniform DTRs $(g_1, g_2) = (0,1), (1,1)$, which satisfy the monotonicity constraint. The simultaneous method leads to a higher welfare gain than the uniform DTRs. The backward method results in a lower welfare gain, possibly because the first-best rule is not included in the policy class $\mathcal{G}_2$ due to the additional constraint from the optimal starting problem.

Table 3: Estimated welfare gains for the start-time decision problem

	Share of population to be treated		
Dynamic treatment regime	1st stage	2nd stage	Estimated welfare gain
$(g_1, g_2) = (0,1)$	0	1	3.94
$(g_1, g_2) = (1,1)$	1	1	0.80
$(\hat{g}_1^B, \hat{g}_2^B)$	0.0	0.51	1.79
$(\hat{g}_1^S, \hat{g}_2^S)$	0.63	1.0	5.56

Notes: The standard deviation of Y_2 in the sample is 39.65. We use the two-fold sample splitting with a fixed seed. The training sample is used to estimate the DTRs $\hat{g}^B$ and $\hat{g}^S$. The test sample is used to estimate shares of population to be treated and welfare gains of $\hat{g}^B$ and $\hat{g}^S$. The welfare gains of the uniform policies are estimated with the whole sample.

8 Conclusion

This study proposes empirical methods to estimate the optimal DTR over a pre-specified class of feasible DTRs based on the EWM approach. We proposed two estimation methods, the backward DEWM and simultaneous DEWM methods, which estimate the optimal DTR through backward induction and simultaneous maximization, respectively. The former is computationally efficient, but may not consistently estimate the optimal DTR

when the class of feasible DTRs does not include the first-best rule at all stages except for the first stage. Conversely, the latter method can consistently estimate the optimal DTR irrespective of the feasibility of the first-best rule, though it is computationally less efficient. These methods can accommodate exogenous constraints on the class of DTRs and specify different types of dynamic treatment choice problems. We show that each method can achieve the optimal $n^{-1/2}$ rate of convergence of the regret in the experimental data setting. We also modify the simultaneous DEWM to accommodate intertemporal budget/capacity constraints.

Appendix

A Proof of Theorem 3.6

This appendix presents the proof of Theorem 3.6 along with some auxiliary lemmas. Let $\mathrm{SG}(\mathcal{F}) \equiv \{\mathrm{SG}(f) : f \in \mathcal{F}\}$ be a collection of subgraphs over a class of functions $\mathcal{F}$, where the subgraph of a real-valued function $f : \mathcal{Z} \to \mathbb{R}$ is defined as the set $\mathrm{SG}(f) \equiv \{(z,t) \in \mathcal{Z} \times \mathbb{R} : t \leq f(z)\}$. We consider the VC-dimension of $\mathrm{SG}(\mathcal{F})$ as a complexity measure of $\mathcal{F}$, where its definition is given in Supplemental Appendix F.

The following lemma establishes the link between the VC-dimension of a class of feasible DTRs and the VC-dimension of a class of subgraphs of functions on $\mathcal{Z}$.

Lemma A.1. *Suppose that Assumption 2.3 holds. Let $r : \mathcal{Z} \to \mathbb{R}$ be any function. For any integers s and t with $1 \leq s \leq t \leq T$, a class of functions from $\mathcal{Z}$ to $\mathbb{R}$*

$$\mathcal{F}_{s:t} \equiv \{f(z) = 1\{g_s(h_s) = d_s, \ldots, g_t(h_t) = d_t\} \cdot r(z) : (g_s, \ldots, g_t) \in \mathcal{G}_s \times \cdots \times \mathcal{G}_t\}$$

is a VC-subgraph class of functions with $VC(SG(\mathcal{F}_{s:t})) \leq \sum_{j=s}^{t} v_j$.

Proof. The proof is presented in Supplemental Appendix F. □

The next lemma, which corresponds to Lemma A.4 of Kitagawa and Tetenov (2018b), gives a uniform upper bound for the mean of a supremum of centered empirical processes indexed by a VC-subgraph class of functions. This is a fundamental result in the literature on empirical process theory and its proof can be found, for example, in van der Vaart and Wellner (1996) and Kitagawa and Tetenov (2018b).

Lemma A.2. *(Lemma A.4 in Kitagawa and Tetenov (2018b)) Let $\mathcal{F}$ be a class of uniformly bounded functions on $\mathcal{Z}$, 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 of functions with VC-dimension $v < \infty$. Then there is a universal constant C such that*

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

holds for all $n \geq 1$.

Before proceeding to the proofs of the main theorems, we define

$$\begin{aligned}
\tilde{Q}_t\left(g_t, \ldots, g_T\right) &\equiv E_P\left[q_t\left(Z, g_t, \ldots, g_T\right)\right] \\
&= E_P\left[\sum_{s=t}^{T}\left\{\frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = g_\ell\left(H_\ell\right)\right\}\right)\gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell\left(D_\ell, H_\ell\right)}\right\}\right], \qquad (16) \\
\tilde{Q}_{nt}\left(g_t, \ldots, g_T\right) &\equiv E_n\left[q_t\left(Z, g_t, \ldots, g_T\right)\right] \\
&= E_n\left[\sum_{s=t}^{T}\left\{\frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = g_\ell\left(H_\ell\right)\right\}\right)\gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell\left(D_\ell, H_\ell\right)}\right\}\right].
\end{aligned}$$

We further define

$$\begin{aligned}
\Delta\tilde{Q}_t &\equiv \tilde{Q}_t\left(g_t^*, \ldots, g_T^*\right) - \tilde{Q}_t\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right), \\
\Delta\tilde{Q}_t^\dagger &\equiv \tilde{Q}_t\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_t\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right).
\end{aligned}$$

The following lemma will be used in the proof of Theorem 3.6 (ii) for the backward DEWM method.

Lemma A.3. *Suppose that Assumptions 2.1, 2.4, and 3.1 hold for a pair $(P, \mathcal{G})$. Then the following hold: (i) for any $t = 1, \ldots, T-1$ and $s = t+1, \ldots, T$,*

$$\tilde{Q}_t\left(g_t^*, \ldots, g_T^*\right) - \tilde{Q}_t\left(g_t^*, \ldots, g_s^*, \hat{g}_{s+1}^B, \ldots, \hat{g}_T^B\right) \leq \frac{1}{\prod_{\ell=t}^{s}\kappa_\ell}\Delta\tilde{Q}_{s+1};$$

(ii)

$$\Delta\tilde{Q}_1 \leq \Delta\tilde{Q}_1^\dagger + \sum_{s=1}^{T-1}\frac{2^{s-1}}{\prod_{t=1}^{s}\kappa_t}\Delta\tilde{Q}_{s+1}^\dagger.$$

Proof. (i) Let $\tilde{Q}_t\left(g_t, \ldots, g_T; h_t\right) \equiv E_P\left[q_t\left(Z, g_t, \ldots, g_T\right)|H_t = h_t\right]$. For any integers s and t such that $1 \leq t < s \leq T$, it follows that

$$\begin{aligned}
&\tilde{Q}_t\left(g_t^*, \ldots, g_T^*\right) - \tilde{Q}_t\left(g_t^*, \ldots, g_s^*, \hat{g}_{s+1}^B, \ldots, \hat{g}_T^B\right) \\
&= E_P\left[\frac{\prod_{\ell=t}^{s} 1\left\{D_\ell = g_\ell^*\left(H_\ell\right)\right\}}{\prod_{\ell=t}^{s} e_\ell\left(D_\ell, H_\ell\right)}\left(\tilde{Q}_{s+1}\left(g_{s+1}^*, \ldots, g_T^*; H_{s+1}\right) - \tilde{Q}_{s+1}\left(\hat{g}_{s+1}^B, \ldots, \hat{g}_T^B; H_{s+1}\right)\right)\right] \\
&\leq \frac{1}{\prod_{\ell=t}^{s}\kappa_\ell}E_P\left[\tilde{Q}_{s+1}\left(g_{s+1}^*, \ldots, g_T^*; H_{s+1}\right) - \tilde{Q}_{s+1}\left(\hat{g}_{s+1}^B, \ldots, \hat{g}_T^B; H_{s+1}\right)\right]
\end{aligned}$$

$$= \frac{1}{\prod_{\ell=t}^{s} \kappa_\ell} \Delta \tilde{Q}_{s+1},$$

where the first equality follows from Assumption 2.1 and the inequality follows from Assumption 2.4 and because $\tilde{Q}_{s+1}\left(g_{s+1}^*, \ldots, g_T^*; H_{s+1}\right) - \tilde{Q}_{s+1}\left(\hat{g}_{s+1}^B, \ldots, \hat{g}_T^B; H_{s+1}\right) \geq 0$ holds a.s. under Assumptions 2.1 and 3.1.

(ii) Note that

$$\Delta \tilde{Q}_T = \tilde{Q}_T\left(g_T^*\right) - \tilde{Q}_T\left(\hat{g}_T^B\right) = \Delta \tilde{Q}_T^\dagger.$$

Then, for $t = T-1$, we have

$$\begin{aligned}
\Delta \tilde{Q}_{T-1} &= \tilde{Q}_{T-1}\left(g_{T-1}^*, g_T^*\right) - \tilde{Q}_{T-1}\left(\hat{g}_{T-1}^B, \hat{g}_T^B\right) \\
&= \tilde{Q}_{T-1}\left(g_{T-1}^*, g_T^*\right) - \tilde{Q}_{T-1}\left(g_{T-1}^*, \hat{g}_T^B\right) + \tilde{Q}_{T-1}\left(g_{T-1}^*, \hat{g}_T^B\right) - \tilde{Q}_{T-1}\left(\hat{g}_{T-1}^B, \hat{g}_T^B\right) \\
&\leq \frac{1}{\kappa_{T-1}} \Delta \tilde{Q}_T^\dagger + \Delta \tilde{Q}_{T-1}^\dagger,
\end{aligned}$$

where the inequality follows from Lemma A.3 (i).

Generally, for any $k = 1, \ldots, T-1$, it follows that

$$\begin{aligned}
\Delta \tilde{Q}_{T-k} &= \tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_T^*\right) - \tilde{Q}_{T-k}\left(\hat{g}_{T-k}^B, \ldots, \hat{g}_T^B\right) \\
&= \sum_{s=T-k}^{T} \left[\tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_s^*, \hat{g}_{s+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_{s-1}^*, \hat{g}_s^B, \ldots, \hat{g}_T^B\right)\right] \\
&\leq \sum_{s=T-k}^{T} \left[\tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_T^*\right) - \tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_{s-1}^*, \hat{g}_s^B, \ldots, \hat{g}_T^B\right)\right] \\
&= \sum_{s=T-k+1}^{T} \left[\tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_T^*\right) - \tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_{s-1}^*, \hat{g}_s^B, \ldots, \hat{g}_T^B\right)\right] + \Delta \tilde{Q}_{T-k}^\dagger \\
&\leq \sum_{s=T-k+1}^{T} \frac{1}{\prod_{\ell=T-k}^{s-1} \kappa_\ell} \Delta \tilde{Q}_s + \Delta \tilde{Q}_{T-k}^\dagger,
\end{aligned}$$

where the second line follows by taking a telescope sum; the third line follows from the fact that $\left(g_{s+1}^*, \ldots, g_T^*\right)$ maximizes $\tilde{Q}_{T-k}\left(g_{T-k}^*, \ldots, g_s^*, \cdot, \ldots, \cdot\right)$ over $\mathcal{G}_{s+1} \times \cdots \times \mathcal{G}_T$ under Assumption 3.1; the last line follows from Lemma A.3 (i).

Then, recursively, the following hold:

$$\begin{aligned}
\Delta\tilde{Q}_{T-1} &\leq \frac{1}{\kappa_{T-1}}\Delta\tilde{Q}_T + \Delta\tilde{Q}^{\dagger}_{T-1} = \frac{1}{\kappa_{T-1}}\Delta\tilde{Q}^{\dagger}_T + \Delta\tilde{Q}^{\dagger}_{T-1}, \\
\Delta\tilde{Q}_{T-2} &\leq \frac{1}{\kappa_{T-2}}\Delta\tilde{Q}_{T-1} + \frac{1}{\kappa_{T-2}\kappa_{T-1}}\Delta\tilde{Q}_T + \Delta\tilde{Q}^{\dagger}_{T-2} \\
&\leq \frac{2}{\kappa_{T-2}\kappa_{T-1}}\Delta\tilde{Q}^{\dagger}_T + \frac{1}{\kappa_{T-2}}\Delta\tilde{Q}^{\dagger}_{T-1} + \Delta\tilde{Q}^{\dagger}_{T-2}, \\
&\vdots \\
\Delta\tilde{Q}_{T-k} &\leq \sum_{s=1}^{k} \frac{2^{k-s}}{\prod_{t=T-k}^{T-s}\kappa_t}\Delta\tilde{Q}^{\dagger}_{T-s+1} + \Delta\tilde{Q}^{\dagger}_{T-k}.
\end{aligned}$$

Therefore, when $k = T-1$, we have

$$\begin{aligned}
\Delta\tilde{Q}_1 &\leq \Delta\tilde{Q}^{\dagger}_1 + \sum_{s=1}^{T-1} \frac{2^{T-1-s}}{\prod_{t=1}^{T-s}\kappa_t}\Delta\tilde{Q}^{\dagger}_{T-s+1} \\
&= \Delta\tilde{Q}^{\dagger}_1 + \sum_{s=1}^{T-1} \frac{2^{s-1}}{\prod_{t=1}^{s}\kappa_t}\Delta\tilde{Q}^{\dagger}_{s+1}.
\end{aligned}$$

□

We are now prepared to give the proof of Theorem 3.6. We first give the proof for the simultaneous DEWM method.

Proof of Theorem 3.6 (i). Let $P \in \mathcal{P}(M, \kappa, \mathcal{G})$ be fixed. Define $W_t(\underline{g}_t) \equiv E_P\left[\gamma_t \widetilde{Y}_t\left(\underline{g}_t\right)\right]$. Note that $W_t(\underline{g}_t) = E_P\left[w_t^S(Z, \underline{g}_t)\right]$ holds under Assumption 2.1, where $w_t^S(Z, \underline{g}_t)$ is defined in Section 3.2. Note also that $W(g) = \sum_{t=1}^{T} W_t(\underline{g}_t)$. Let $W_{nt}(\underline{g}_t)$ and $W_n(g)$ be defined as $W_{nt}(\underline{g}_t) \equiv \frac{1}{n}\sum_{i=1}^{n} w_t^S(Z_i, \underline{g}_t)$ and $W_n(g) \equiv \sum_{t=1}^{T} W_{nt}(\underline{g}_t)$, respectively.

It follows, for any $g \in \mathcal{G}$, that

$$\begin{aligned}
E_{P^n}\left[W(g) - W\left(\hat{g}^S\right)\right] &= E_{P^n}\left[W(g) - W_n(g)\right] + E_{P^n}\left[W_n(g) - W\left(\hat{g}^S\right)\right] \\
&\leq E_{P^n}\left[W(g) - W_n(g)\right] + E_{P^n}\left[W_n\left(\hat{g}^S\right) - W\left(\hat{g}^S\right)\right] \\
&\leq 2E_{P^n}\left[\sup_{g\in\mathcal{G}}\left|W_n(g) - W(g)\right|\right] \\
&= 2E_{P^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{t=1}^{T}\left(W_{nt}(\underline{g}_t) - W_t(\underline{g}_t)\right)\right|\right]
\end{aligned}$$

$$\leq 2\sum_{t=1}^{T} E_{P^n}\left[\sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} \left|W_{nt}(\underline{g}_t) - W_t(\underline{g}_t)\right|\right], \tag{17}$$

where the second line follows from the fact that $\hat{g}^S$ maximizes $W_n(\cdot)$ over $\mathcal{G}$, and the fourth line follows from the definition of $W_n(\cdot)$ and equation (2).

Applying Lemma A.2, combined with Lemma A.1, to each term in (17) leads to the following: for each $t = 1, \ldots, T$,

$$E_{P^n}\left[\sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} \left|W_{nt}(\underline{g}_t) - W_t(\underline{g}_t)\right|\right] \leq C \frac{\gamma_t M_t / 2}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}},$$

where C is the same universal constant that appears in Lemma A.2. Combining this result with (17), we obtain

$$E_{P^n}\left[W_{\mathcal{G}}^* - W\left(\hat{g}^S\right)\right] \leq C \sum_{t=1}^{T} \left\{ \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}} \right\}.$$

Since this upper bound does not depend on $P \in \mathcal{P}(M, \kappa, \mathcal{G})$, the upper bound is uniform over $\mathcal{P}(M, \kappa, \mathcal{G})$. □

We next present the proof for the backward DEWM method.

Proof of Theorem 3.6 (ii). Let $P \in \mathcal{P}(M, \kappa, \mathcal{G})$ be fixed. Let g^* be defined in Section 3.1. It follows under Assumptions 2.1 and 3.1 that

$$W_{\mathcal{G}}^* - W\left(\hat{g}^B\right) = \tilde{Q}_1\left(g^*\right) - \tilde{Q}_1\left(\hat{g}^B\right) \leq \Delta\tilde{Q}_1.$$

Then, from Lemma A.3 (ii),

$$W_{\mathcal{G}}^* - W\left(\hat{g}^B\right) \leq \Delta\tilde{Q}_1^{\dagger} + \sum_{s=1}^{T-1} \frac{2^{s-1}}{\prod_{t=1}^{s} \kappa_t} \Delta\tilde{Q}_{s+1}^{\dagger}.$$

Thus, we have

$$E_{P^n}\left[W_{\mathcal{G}}^* - W\left(\hat{g}^B\right)\right] \leq E_{P^n}\left[\Delta\tilde{Q}_1^{\dagger}\right] + \sum_{s=1}^{T-1} \frac{2^{s-1}}{\prod_{t=1}^{s} \kappa_t} E_{P^n}\left[\Delta\tilde{Q}_{s+1}^{\dagger}\right]. \tag{18}$$

Regarding $\Delta\tilde{Q}_t^{\dagger}$ for each t, it follows that

$$\begin{aligned}
\Delta\tilde{Q}_t^{\dagger} &= \tilde{Q}_t\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_t\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right) \\
&= \tilde{Q}_t\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_{nt}\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) + \tilde{Q}_{nt}\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_t\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right) \\
&\leq \tilde{Q}_t\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_{nt}\left(g_t^*, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right) + \tilde{Q}_{nt}\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right) - \tilde{Q}_t\left(\hat{g}_t^B, \ldots, \hat{g}_T^B\right) \\
&\leq 2 \sup_{(g_t,\ldots,g_T)\in\mathcal{G}_t\times\cdots\times\mathcal{G}_T} \left|\tilde{Q}_{nt}\left(g_t, \ldots, g_T\right) - \tilde{Q}_t\left(g_t, \ldots, g_T\right)\right|, \qquad (19)
\end{aligned}$$

where the first inequality follows from the fact that $\hat{g}_t^B$ maximizes $\tilde{Q}_{nt}\left(\cdot, \hat{g}_{t+1}^B, \ldots, \hat{g}_T^B\right)$ over $\mathcal{G}_t$. Because $\left\|\tilde{Q}_t\left(g_t, \ldots, g_T\right)\right\|_{\infty} \leq \sum_{s=t}^{T}(\gamma_s M_s/2)/\left(\prod_{\ell=t}^{s}\kappa_\ell\right)$ holds under Assumptions 2.2 and 2.4, by applying Lemmas A.1 and A.2 to the following class of functions:

$$\left\{\sum_{s=t}^{T}\left\{\frac{\left(\prod_{\ell=t}^{s} 1\left\{D_\ell = g_\ell\left(H_\ell\right)\right\}\right)\gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell\left(D_\ell, H_\ell\right)}\right\} : \left(g_t, \ldots, g_T\right) \in \mathcal{G}_t \times \cdots \times \mathcal{G}_T\right\},$$

we have

$$E_{P^n}\left[\sup_{(g_t,\ldots,g_T)\in\mathcal{G}_t\times\cdots\times\mathcal{G}_T}\left|\tilde{Q}_{nt}\left(g_t, \ldots, g_T\right) - \tilde{Q}_t\left(g_t, \ldots, g_T\right)\right|\right] \leq C\left(\sum_{s=t}^{T}\frac{\gamma_s M_s/2}{\prod_{\ell=t}^{s}\kappa_\ell}\right)\sqrt{\frac{\sum_{s=t}^{T} v_s}{n}}.$$

Combining this with equations (18) and (19) leads to

$$\begin{aligned}
E_{P^n}\left[W_{\mathcal{G}}^* - W\left(\hat{g}^B\right)\right] \leq C\sum_{t=1}^{T}\left\{\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\sqrt{\frac{\sum_{s=1}^{t} v_s}{n}}\right\} \\
+ \sum_{t=2}^{T}\frac{2^{t-2}}{\prod_{s=1}^{t-1}\kappa_s}\left(C\sum_{s=t}^{T}\left\{\frac{\gamma_s M_s}{\prod_{\ell=t}^{s}\kappa_\ell}\sqrt{\frac{\sum_{\ell=t}^{s} v_\ell}{n}}\right\}\right),
\end{aligned}$$

where C is the same universal constant that appears in Lemma A.2. Since this upper bound does not depend on $P \in \mathcal{P}\left(M, \kappa, \mathcal{G}\right)$, the upper bound is uniform over $\mathcal{P}\left(M, \kappa, \mathcal{G}\right)$.

□

B Proof of Theorem 4.1

This appendix presents the proof of Theorem 4.1. We first introduce several lemmas that will be used in the proof of Theorem 4.1.

The following lemma provides a concentration inequality that is frequently used in the literature on statistical learning theory, the proof of which can be found, for example, in Mohri et al. (2012).

Lemma B.1. *(McDiarmid's Inequality): Let $S = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ 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:*

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

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

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

The following lemma gives a finite-sample upper bound on $\sup_{g \in \mathcal{G}} |W(g) - W_n(g)|$ that holds with a high probability.

Lemma B.2. *Suppose that the underlying distribution P satisfies Assumptions 2.1, 2.2, and 2.4 and that $\mathcal{G}$ satisfies Assumption 2.3. Then, for any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:*

$$\sup_{g \in \mathcal{G}} |W(g) - W_n(g)| \leq \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[\frac{\gamma_t M_t}{\prod_{s=1}^t \kappa_s} \cdot \left(C \sqrt{\sum_{s=1}^t v_s} + \sqrt{\frac{\log(1/\delta)}{2}} \right) \right]. \tag{20}$$

Proof. The proof follows a similar argument as that of Corollary 3.4 of Mohri et al. (2012). We will evaluate $\sup_{g \in \mathcal{G}} |W(g) - W_n(g)|$. Let $S = (Z_1, \ldots, Z_n)$ be the sample and define $A(S) \equiv \sup_{g \in \mathcal{G}} \{W(g) - W_S(g)\}$, where, for any sample S with size n, $W_S(g)$ is defined as $W_n(g)$ that uses the sample S.

Introduce $S' = (Z_1, \ldots, Z_{n-1}, Z_n')$, an i.i.d. sample that is different from S with respect

to the final component. Then, it follows that

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

where the last inequality follows from the fact that under Assumptions 2.2 and 2.4, $w_t^S \left(Z_i, \underline{g}_t\right)$ is bounded from above by $(\gamma_t M_t/2) / \left(\prod_{s=1}^{t} \kappa_s\right)$.

Since we have

$$|A(S) - A(S')| \leq \frac{1}{n} \sum_{t=1}^{T} \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s},$$

applying Lemma B.1 leads to

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

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

$$P\left(|A(S) - E_{P^n}[A(S)]| \leq \left(\sum_{t=1}^{T} \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \sqrt{\frac{\log(1/\delta)}{2n}} \right) \geq 1 - \delta. \tag{21}$$

Subsequently, we will evaluate $E_{P^n}[A(S)]$. Since

$$\begin{aligned} E_{P^n}[A(S)] &= E_{P^n}\left[\sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{T} \left(W_{nt}(\underline{g}_t) - W_t(\underline{g}_t) \right) \right| \right] \\ &\leq \sum_{t=1}^{T} E_{P^n}\left[\sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} \left| W_{nt}(\underline{g}_t) - W_t(\underline{g}_t) \right| \right], \end{aligned}$$

applying Lemma A.2 combined with Lemma A.1 leads to

$$E_{P^n}[A(S)] \leq C \sum_{t=1}^{T} \left\{ \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}} \right\}, \tag{22}$$

where C is the same constant that appears in Lemma A.2.

Consequently, combining (21) and (22), for any $\delta \in (0,1)$, it follows with probability at least $1-\delta$ that

$$\begin{aligned} \sup_{g \in \mathcal{G}} |W(g) - W_n(g)| &\leq C \sum_{t=1}^{T} \left[\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \sqrt{\frac{\sum_{s=1}^{t} v_s}{n}} \right] + \left(\sum_{t=1}^{T} \frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \right) \sqrt{\frac{\log(1/\delta)}{2n}} \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \cdot \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log(1/\delta)}{2}} \right) \right]. \end{aligned}$$

□

The following lemma shows that a class of feasible DTRs that satisfy the empirical budget/capacity constraints (13) contains the optimal DTR with high probability.

Lemma B.3. *Suppose that the underlying distribution P satisfies Assumption 2.1 and that $\sum_{t=1}^{T} K_{tb} = 1$ holds for all $b = 1, \ldots, B$. For $k > 0$, let $\tilde{g}^* = (\tilde{g}_1^*, \ldots, \tilde{g}_T^*)$ be a solution of the constrained maximization problem (11) with C_b replaced by $C_b - k + \alpha_n$, where we suppose that such a solution exists. Define*

$$\mathcal{G}_{\alpha_n}^S \equiv \left\{ g \in \mathcal{G} : \sum_{t=1}^{T} K_{tb} \hat{E}\left[g_t\left(\widetilde{H}_t\left(\underline{g}_{t-1} \right) \right) \right] \leq C_b + \alpha_n \text{ for } b = 1, \ldots, B \right\},$$

which is a subset of DTRs that satisfy the sample budget constraints (13). Then, for any $\delta \in (0,1)$, $P\left(\tilde{g}^* \in \mathcal{G}_{\alpha_n}^S\right) \geq 1 - B \cdot \exp(-2nk^2)$ holds.

Proof. It follows that

$$\begin{aligned} P\left(\tilde{g}^* \notin \mathcal{G}_{\alpha_n}^S\right) &= P\left(\max_{b=1,\ldots,B} \left\{ \sum_{t=1}^{T} K_{tb} \hat{E}\left[\tilde{g}_t^*\left(\widetilde{H}_t\left(\underline{\tilde{g}}_{t-1}^* \right) \right) \right] - C_b \right\} > \alpha_n \right) \\ &\leq \sum_{b=1}^{B} P\left(\sum_{t=1}^{T} K_{tb} \hat{E}\left[\tilde{g}_t^*\left(\widetilde{H}_t\left(\underline{\tilde{g}}_{t-1}^* \right) \right) \right] - C_b > \alpha_n \right) \end{aligned}$$

$$\leq \sum_{b=1}^{B} P\left(\sum_{t=1}^{T} K_{tb}\hat{E}\left[\tilde{g}_t^*\left(\widetilde{H}_t\left(\underline{\tilde{g}}_{t-1}^*\right)\right)\right] - \sum_{t=1}^{T} K_{tb}E_P\left[\tilde{g}_t^*\left(\widetilde{H}_t\left(\underline{\tilde{g}}_{t-1}^*\right)\right)\right] > k\right),$$

where the second inequality follows from the fact that $\tilde{g}^*$ satisfies the population budget/capacity constraints (10) with C_b replaced by $C_b - k + \alpha_n$.

By Hoeffding's inequality, it follows for each $b = 1, \ldots, B$ that

$$P\left(\sum_{t=1}^{T} K_{tb}\hat{E}\left[\tilde{g}_t^*\left(\widetilde{H}_t\left(\underline{\tilde{g}}_{t-1}^*\right)\right)\right] - \sum_{t=1}^{T} K_{tb}E_P\left[g_t^*\left(\widetilde{H}_t\left(\underline{g}_{t-1}^*\right)\right)\right] > k\right)$$
$$\leq \exp\left\{-\frac{2nk^2}{\left(\sum_{t=1}^{T} K_{tb}\right)^2}\right\} = \exp\left(-2nk^2\right),$$

where the equality follows from the scale normalization $\sum_{t=1}^{T} K_{tb} = 1$. Thus, we have $P\left(\tilde{g}^* \notin \mathcal{G}_{\alpha_n}^S\right) \leq B \cdot \exp\left(-2nk^2\right)$. Therefore, $P\left(\tilde{g}^* \in \mathcal{G}_{\alpha_n}^S\right) = 1 - P\left(\tilde{g}^* \notin \mathcal{G}_{\alpha_n}^S\right) \geq 1 - B \cdot \exp\left(-2nk^2\right)$. □

Proof of Theorem 4.1 (i). We use the notation $A \leq_\delta B$ to denote that $A \leq B$ holds with probability at least $1 - \delta$. Let $g_{\alpha_n}^*$ be a solution of the constrained maximization problem (11) with C_b replaced by $C_b - k_{(B,n,\delta)} + \alpha_n$.

By Lemma B.3, we have $P\left(g_{\alpha_n}^* \in \mathcal{G}_{\alpha_n}^S\right) \geq 1 - \delta/6$. Thus, $W_n\left(g_{\alpha_n}^*\right) \leq_{\delta/6} W_n\left(\hat{g}^{bdgt}\right)$ holds because $\hat{g}^{bdgt}$ maximizes $W_n(\cdot)$ over $\mathcal{G}_{\alpha_n}^S$. Note that $W\left(g_{\alpha_n}^*\right) = W_{\mathcal{G}}^{*,bdgt}$. By combining the fact that $W_n\left(g_{\alpha_n}^*\right) \leq_{\delta/6} W_n\left(\hat{g}^{bdgt}\right)$ with (20), it follows that

$$\begin{aligned}
W_{\mathcal{G}}^{*,bdgt} &= W\left(g_{\alpha_n}^*\right) \\
&\leq_{\delta/6} W_n\left(g_{\alpha_n}^*\right) + \frac{1}{\sqrt{n}}\sum_{t=1}^{T}\left[\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\cdot\left(C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log(6/\delta)}{2}}\right)\right] \\
&\leq_{\delta/6} W_n\left(\hat{g}^{bdgt}\right) + \frac{1}{\sqrt{n}}\sum_{t=1}^{T}\left[\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\cdot\left(C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log(6/\delta)}{2}}\right)\right] \\
&\leq_{\delta/6} W\left(\hat{g}^{bdgt}\right) + \frac{2}{\sqrt{n}}\sum_{t=1}^{T}\left[\frac{\gamma_t M_t}{\prod_{s=1}^{t}\kappa_s}\cdot\left(C\sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log(6/\delta)}{2}}\right)\right].
\end{aligned}$$

The first inequality follows from the inequality in (20); the second inequality follows from the fact that $\hat{g}^{bdgt}$ maximizes $W_n(\cdot)$ over $\mathcal{G}_{\alpha_n}^S$ and $g_{\alpha_n}^* \in \mathcal{G}_{\alpha_n}^S$ holds with probability at

least $1-\delta/6$; the third inequality follows from the inequality in (20). Overall, we have

$$W_{\mathcal{G}}^{*,bdgt} \leq_{\delta/2} W\left(\hat{g}^{bdgt}\right) + \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[\frac{\gamma_t M_t}{\prod_{s=1}^{t} \kappa_s} \cdot \left(2C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{2 \log\left(6/\delta\right)} \right) \right]. \tag{23}$$

Applying the same argument as in the proof of Lemma B.2, it follows for each $b = 1, \ldots, B$ that

$$\begin{aligned}
&\left| E_n \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^{bdgt} \left(\widetilde{H}_t \left(\underline{\hat{g}}_{t-1}^{bgdt} \right) \right) \right] - E_P \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^{bdgt} \left(\widetilde{H}_t \left(\underline{\hat{g}}_{t-1}^{bgdt} \right) \right) \right] \right| \\
\leq & \sum_{t=1}^{T} \sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} \left| E_n \left[K_{tb} g_t \left(\widetilde{H}_t \left(\underline{g}_{t-1} \right) \right) \right] - E_P \left[K_{tb} g_t \left(\widetilde{H}_t \left(\underline{g}_{t-1} \right) \right) \right] \right| \\
\leq_{\delta} & \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[K_{tb} \cdot \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log\left(1/\delta\right)}{2}} \right) \right].
\end{aligned} \tag{24}$$

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

$$\begin{aligned}
E_P \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^{bdgt} \left(\widetilde{H}_t \left(\underline{\hat{g}}_{t-1}^{bgdt} \right) \right) \right] \leq_{\delta/(2B)} & E_n \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^{bdgt} \left(\widetilde{H}_t \left(\underline{\hat{g}}_{t-1}^{bgdt} \right) \right) \right] \\
& + \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[K_{tb} \cdot \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log\left(2B/\delta\right)}{2}} \right) \right] \\
& \leq C_b + \alpha_n + \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[K_{tb} \cdot \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log\left(2B/\delta\right)}{2}} \right) \right],
\end{aligned}$$

where the first inequality follows from the inequality in (24), and the second inequality follows from the fact that $\hat{g}^{bdgt} \in \mathcal{G}_{\alpha_n}^{S}$. Thus, the following holds with probability at least $1-\delta$: for any $b \in \{1, \ldots, B\}$,

$$\begin{aligned}
& E_P \left[\sum_{t=1}^{T} K_{tb} \hat{g}_t^{bdgt} \left(\widetilde{H}_t \left(\underline{\hat{g}}_{t-1}^{bgdt} \right) \right) - C_b \right] \\
\leq_{\delta/(2B)} & \alpha_n + \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \left[K_{tb} \cdot \left(C \sqrt{\sum_{s=1}^{t} v_s} + \sqrt{\frac{\log\left(2B/\delta\right)}{2}} \right) \right].
\end{aligned} \tag{25}$$

The result follows from combining the probability inequalities (23) and (25) for all

$b = 1, \ldots, B$. □

References

Aaberge, R., T. Havnes, and M. Mogstad (2013): "A theory for ranking distribution functions," *Available at SSRN 2363225*.

Achilles, C., H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word (2008): "Tennessee's Student Teacher Achievement Ratio (STAR) project," *Harvard Dataverse*, https://doi.org/10.7910/DVN/SIWH9F.

Angrist, J. (2009): "MHE Data Archive," URL:https://economics.mit.edu/people/faculty/josh–angrist/mhe–data–archive, accessed on October 31, 2024.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Athey, S. and S. Wager (2021): "Policy learning with observational data," *Econometrica*, 89, 133–161.

Bhattacharya, D. and P. Dupas (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167, 168–196.

Blackorby, C. and D. Donaldson (1978): "Measures of relative equality and their meaning in terms of social welfare," *Journal of Economic Theory*, 18, 59–80.

Chakraborty, B. and E. E. M. Moodie (2013): *Statistical Methods for Dynamic Treatment Regimes*, New York: Springer.

Chakraborty, B. and S. A. Murphy (2014): "Dynamic treatment regimes," *Annual Review of Statistics and Its Application*, 1, 447–464.

Chamberlain, G. (2012): "Bayesian aspects of treatment choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford: Oxford University Press.

Dehejia, R. H. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173.

Donaldson, D. and J. A. Weymark (1980): "A single-parameter generalization of the Gini indices of inequality," *Journal of economic Theory*, 22, 67–86.

——— (1983): "Ethically flexible Gini indices for income distributions in the continuum," *Journal of Economic Theory*, 29, 353–358.

Goel, K., C. Dann, and E. Brunskill (2017): "Sample efficient policy search for optimal stopping domains," *arXiv preprint arXiv:1702.06238*.

Han, S. (2021): "Identification in nonparametric models for dynamic treatment effects," *Journal of Econometrics*, 225, 132–147.

——— (2023): "Optimal dynamic treatment regimes and partial welfare ordering," *Journal of the American Statistical Association*, 1–11.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2016): "Dynamic treatment effects," *Journal of Econometrics*, 191, 276–292.

Heckman, J. J. and S. Navarro (2007): "Dynamic discrete choice and dynamic treatment effects," *Journal of Econometrics*, 136, 341–396.

Hirano, K. and J. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77, 1683–1701.

Jacka, S. D. (1991): "Optimal stopping and the American put," *Mathematical Finance*, 1, 1–14.

Kallus, N. (2021): "More efficient policy learning via optimal retargeting," *Journal of the American Statistical Association*, 116, 646–658.

Kitagawa, T., S. Sakaguchi, and A. Tetenov (2021): "Constrained classification and policy learning," *arXiv preprint arXiv:2106.12886*.

Kitagawa, T. and A. Tetenov (2018a): "Supplement to "Who should be treated? Empirical welfare maximization methods for treatment choice"," Econometrica Supplemental Material, 86.

——— (2018b): "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, 86, 591–616.

——— (2021): "Equality-minded treatment choice," *Journal of Business & Economic Statistics*, 39, 561–574.

Kock, A. B. and M. Thyrsgaard (2018): "Optimal sequential treatment allocation," *arXiv preprint arXiv:1705.09952*.

Kolsrud, J., C. Landais, P. Nilsson, and J. Spinnewijn (2018): "The optimal timing of unemployment benefits: Theory and evidence from Sweden," *American Economic Review*, 108, 985–1033.

Krueger, A. B. (1999): "Experimental estimates of education production functions," *Quarterly Journal of Economics*, 114, 497–532.

Laber, E. B., D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy (2014): "Dynamic treatment regimes: Technical challenges and applications," *Electronic Journal of Statistics*, 8, 1225–1272.

Laha, N., A. Sonabend-W, R. Mukherjee, and T. Cai (2024): "Finding the optimal dynamic treatment regimes using smooth Fisher consistent surrogate loss," *The Annals of Statistics*, 52, 679–707.

Lechner, M. (2009): "Sequential causal models for the evaluation of labor market programs," *Journal of Business & Economic Statistics*, 27, 71–83.

Manski, C. F. (2004): "Statistical treatment rules for heterogeneous populations," *Econometrica*, 72, 1221–1246.

Mbakop, E. and M. Tabord-Meehan (2021): "Model selection for treatment choice: Penalized welfare maximization," *Econometrica*, 89, 825–848.

Meyer, B. D. (1995): "Lessons from the U.S. unemployment insurance experiments," *Journal of Economic Literature*, 33, 91–131.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012): *Foundations of Machine Learning*, Cambridge, MA: MIT Press.

Moodie, E., B. Chakraborty, and M. S. Kramer (2012): "Q-learning for estimating optimal dynamic treatment rules from observational data," *Canadian Journal of Statistics*, 40, 629–645.

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

——— (2005): "A generalization error for Q-learning." *Journal of Machine Learning Research*, 6, 1073–1097.

Nie, X., E. Brunskill, and S. Wager (2021): "Learning when-to-treat policies," *Journal of the American Statistical Association*, 116, 392–409.

Robins, J. M. (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.

——— (1989): "The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies," in *Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, Washington D.C.: U.S. Public Health Service, National Center for Health Services Research, 113–159.

——— (1997): "Causal inference from complex longitudinal data in latent variable modeling and applications to causality," in *Lecture Notes in Statistics*, ed. by M. Berkane, New York: Springer, 69–117.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn (1992): "G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients," *Epidemiology*, 3, 319–336.

Rodríguez, J., F. Saltiel, and S. Urzúa (2022): "Dynamic treatment effects of job training," *Journal of Applied Econometrics*, 37, 242–269.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica*, 999–1033.

Sakaguchi, S. (2024): "Robust learning for optimal dynamic treatment regimes with observational data," ArXiv:2404.00221.

——— (2025): "Supplement to 'Estimation of Optimal Dynamic Treatment Assignment Rules under Policy Constraints'," *Quantitative Economics Supplemental Material*.

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151, 70–81.

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166, 138–156.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166, 157–165.

Tsiatis, A. A., M. Davidian, S. T. Holloway, and E. B. Laber (2019): *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*, CRC press.

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer.

Van Moerbeke, P. (1976): "On optimal stopping and free boundary problems," *Archive for Rational Mechanics and Analysis*, 60, 101–148.

Weymark, J. A. (1981): "Generalized Gini inequality indices," *Mathematical Social Sciences*, 1, 409–430.

Zhao, Y. Q., D. Zeng, E. B. Laber, and M. R. Kosorok (2015): "New statistical learning methods for estimating optimal dynamic treatment regimes," *Journal of the American Statistical Association*, 110, 583–598.

Zhou, Z., S. Athey, and S. Wager (2023): "Offline multi-action policy learning: Generalization and optimization," *Operations Research*, 71, 148–183.