Supplement to "Estimation of Optimal Dynamic Treatment Assignment Rules under Policy Constraints"

Shosei Sakaguchi*

April 8, 2025

Abstract

This document contains Appendices C, D, E, F, G, H, and I of the article "Estimation of Optimal Dynamic Treatment Assignment Rules under Policy Constraints."

Appendix

C Non-additive Welfare Function

In this appendix, we consider a non-additive *social welfare function* (SWF) and provide a simultaneous dynamic EWM approach to estimate the optimal DTR. We consider the equality-minded rank-dependent SWFs introduced by Meyer (1995) and Weymark (1981) and studied by Kitagawa and Tetenov (2021):

$$W_\Lambda(F) \equiv \int_0^\infty \Lambda(F(y))dy, \tag{A.1}$$

where $F(y)$ is the distribution of an outcome and $\Lambda(\cdot) : [0,1] \to [0,1]$ is a non-increasing, non-negative functions with $\Lambda(0) = 1$ and $\Lambda(1) = 0$.

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

An important family of SWFs represented by (A.1) is the extended Gini family (Donaldson and Weymark, 1980, 1983; Aaberge et al., 2013):

$$W_k(F) \equiv \int_0^\infty (1-F(y))^{k-1} dy = \int_0^\infty \Lambda_k\left(F(y)\right) dy$$
$$= \int_0^1 F^{-1}(\tau)\omega_k(\tau) d\tau,$$

where $\Lambda_k(\tau) \equiv (1-\tau)^{k-1}$ and $\omega_k(\tau) \equiv (k-1)(1-\tau)^{k-2}$. The standard Gini social welfare function (Blackorby and Donaldson, 1978; Weymark, 1981) corresponds to the extended Gini social welfare function when $k = 3$, which can also be written as

$$W_{Gini}(F) = E(Y)(1 - I_{Gini}(F)),$$

where $I_{Gini}(F) = 1 - \frac{\int_0^1 F^{-1}(\tau)\cdot 2(1-\tau)d\tau}{E(Y)}$ is the widely used Gini inequality index.

Without loss of generality, we suppose that the target outcome is $\sum_{t=1}^T Y_t$. For any DTR $g = (g_1, \ldots, g_T)$, let $F_g(\cdot)$ denote the distribution of $\sum_{t=1}^T \widetilde{Y}_t(\underline{g}_t)$. We define the rank-dependent SWF of g by

$$W_\Lambda(g) \equiv W_\Lambda(F_g). \tag{A.2}$$

Our goal is to estimate the optimal DTR that maximizes $W_\Lambda(g)$ over the pre-specified class of DTRs $\mathcal{G}$.

We estimate the optimal DTR by simultaneously maximizing the sample analogue of the population welfare function $W_\Lambda(g)$ over $g \in \mathcal{G}$. Let

$$\widehat{F}_g(y) \equiv 1 - \frac{1}{n}\sum_{i=1}^n \left(\frac{\prod_{t=1}^T 1\{D_{it} = g_t(H_{it})\}}{\prod_{t=1}^T e_t(D_{it}, H_{it})} \cdot 1\left\{ \sum_{t=1}^T Y_{it} > y \right\} \right),$$

which is the inverse probability weighting estimator of the distribution of $\sum_{t=1}^T \widetilde{Y}_t(\underline{g}_t)$. The sample analogue of the population welfare $W_\Lambda(g)$ is

$$\widehat{W}_\Lambda(g) \equiv \int_0^\infty \Lambda\left(\widehat{F}_g(y) \vee 0\right) dy,$$

where the maximum ($\vee$) of $\widehat{F}_g(y)$ and 0 is taken because $\widehat{F}_g(y)$ may take values smaller

than 0, for which $\Lambda(\cdot)$ is not defined. The simultaneous DEWM approach estimates the optimal DTR by solving

$$\hat{g}^S \in \arg\max_{g \in \mathcal{G}} \widehat{W}_\Lambda(g).$$

Let $\mathcal{P}$ be a class of distributions of $\left(\underline{A}_T, \{\underline{X}_T(\underline{d}_{T-1})\}_{\underline{d}_{T-1} \in \{0,1\}^{T-1}}, \{\underline{Y}_T(\underline{d}_T)\}_{\underline{d}_T \in \{0,1\}^T}\right)$. The following theorem derives a uniform upper bound of the average welfare loss of $\hat{g}^S$.

Theorem C.1. *Suppose that Assumptions 2.1 and 2.4 hold for any distribution $P \in \mathcal{P}$ and Assumption 2.3 holds for $\mathcal{G}$. Furthermore, suppose that the following hold:*

- *$\Lambda(\cdot) : [0,1] \to [0,1]$ is a non-increasing, convex function with $\Lambda(0) = 1$, $\Lambda(1) = 0$, and its right derivative at 0 is finite;*
- *there exists $\Upsilon < \infty$ such that for all $P \in \mathcal{P}$ and any $\underline{d}_T \in \{0,1\}^T$,*

$$\int_0^\infty \sqrt{P\left(\sum_{t=1}^T Y_t(\underline{d}_t) > y\right)} dy \leq \Upsilon. \tag{A.3}$$

Then the average welfare loss of $\hat{g}^S$ satisfies

$$\sup_{P \in \mathcal{P}} E_{P^n}\left[\sup_{g \in \mathcal{G}} W_\Lambda(g) - W(\hat{g}^S)\right] \leq 2C \mid \Lambda'(0) \mid \frac{\Upsilon}{\prod_{t=1}^T \kappa_t} \sqrt{\frac{\sum_{t=1}^T v_t}{n}} \tag{A.4}$$

for all $n > 1$, where C is a universal constant.

Proof. See Appendix F.4. □

This theorem shows that for a large class of data-generating processes, the rank-dependent SWF of the simultaneous DEWM converges to the optimal welfare no slower than $n^{-1/2}$ rate. This uniform convergence rate of $n^{-1/2}$ coincides with that of the DEWM methods for the linear SWF shown in Theorem 3.6. The convergence rate of $n^{-1/2}$ also coincides with the minimax optimal convergence rate for the rank-dependent SWF in the static treatment case (Theorems 3.1 and 3.2 in Kitagawa and Tetenov (2021)).

D Multiple Treatment

In the main text, we consider the setting of binary treatment assignment for each stage. However, more than a few examples of DTRs involve multiple treatments in practice. In this section, we extend the DEWM to the case of multiple treatment.

Suppose that there are K treatments in each stage. Let $\mathcal{D} \equiv \{1, 2, \ldots, K\}$ denote the treatment space in each stage, and $D_t \in \mathcal{D}$ denote the observed treatment in stage t. Using the same notations as in Section 2, we define the potential outcomes and the potential covariates as $Y_t(\underline{d}_t)$ and $X_t(\underline{d}_{t-1})$, respectively. The observed outcomes and observed covariates are denoted as $Y_t \equiv Y_t(\underline{D}_t)$ and $X_t \equiv X_t(\underline{D}_{t-1})$, respectively. We define the i.i.d. sample $\{Z_i \equiv (D_{it}, X_{it}, Y_{it})_{t=1}^T : i = 1, \ldots, n\}$. Additionally, we define the history H_t and H_{it} in the same manner as described in the main text.

We suppose that the sequential independence assumption holds for multiple treatment.

Assumption D.1. *(Sequential Independence Assumption) For any $t = 1, \ldots, T$ and $\underline{d}_t \in \mathcal{D}^t$,*

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

In the multiple treatment setting, the treatment rule g_t in each stage is a map from $\mathcal{H}_t$ to $\mathcal{D}$. The DTR denoted by g is the sequence $g = (g_1, \ldots, g_T)$. We denote the class of feasible DTRs 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.

In the following subsections, we describe the backward DEWM and simultaneous DEWM for multiple treatment in the experimental data setting.

D.1 Backward DEWM for Multiple Treatments

To guarantee the consistency of the backward estimation procedure, we suppose that the $\mathcal{G}_t$ for $t \geq 2$ contains the first-best rule. As the same way as in the main text, for any $s < t$, we define

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

which is the outcome in stage t that is realized when the treatment assignments from stage 1 to stage $s-1$ are fixed to $\underline{d}_s$ and the subsequent sequential treatment assignment follows $\underline{g}_{(s+1):t}$. We denote $\widetilde{Y}_t\left(\underline{d}_t, \underline{g}_{(t+1):t}\right) = Y_t\left(\underline{d}_t\right)$ when $s = t$.

We suppose that the first-best treatment rule is available for a multiple treatment case in the following sense.

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

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

Then, we can consistently estimate the optimal DTR through the backward induction approach in the same manner as in Section 3.1. Given the propensity scores $\{e_t\left(D_t, H_t\right)\}_{t=1}^T$, let

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

With the backward DEWM, the optimal DTR for multiple treatment is sequentially estimated as follows. In the first step, for the last stage T, the optimal treatment rule in the last stage is estimated as

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

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

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

Throughout this procedure, we obtain the DTR $\hat{g}^B \equiv \left(\hat{g}^B_1, \ldots, \hat{g}^B_T\right)$.

D.2 Simultaneous DEWM for Multiple Treatments

The simultaneous DEWM can also be constructed in the same way as described in Section 3.2. This approach estimates the optimal DTR through the following maximization

problem:

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

with

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

E Doubly Robust Estimation Using Observational Data

We consider doubly robust estimation of the optimal DTR. In Section E.1, we discuss extension of the simultaneous maximization approach to doubly robust approach.[1] However, as discussed in Remark 5.2 in the main text, doubly robust approach with the simultaneous maximization estimation is computationally challenging unless the class of DTR $\mathcal{G}$ is not small. In Section E.2, we construct another doubly-robust simultaneous-maximization approach with computational feasibility under the setting that treatment choice at each stage depends only on the exogenous variables and past treatments.

E.1 Simultaneous Maximization Method

We here consider extending the simultaneous maximization approaches to doubly robust policy learning. Following the doubly robust policy learning of Athey and Wager (2021) and Zhou et al. (2023), we employ cross-fitting to make the estimation of the welfare function and estimation of the optimal DTRs independent; whereby, to reduce the over-fitting. We randomly divide the data set $\{Z_i : i = 1, \ldots, n\}$ into K evenly-sized folds (e.g., $K = 5$). Let I_k be a set of indices of the data in the k-th fold and I_{-k} be a set of

[1] Doubly robust estimators for the optimal DTRs are also proposed by Zhang et al. (2013), Wallace and Moodie (2015), and Ertefaie et al. (2021). Wallace and Moodie (2015) and Ertefaie et al. (2021) propose methods with backward induction, which requires the correct specification of the model for conditional treatment effects at each stage, referred to by Wallace and Moodie (2015) as the blip function. In contrast, the doubly robust approach with simultaneous maximization proposed in my paper does not require the correct specification of the blip function or optimal treatment rules. Zhang et al. (2013) also consider a doubly robust estimation with simultaneous optimization but do not study the theoretical properties of their proposed method.

indices of the data excluded from the k-th fold. Hereafter, for any statistics $\hat{f}$, we denote by $\hat{f}^{-k}$ the corresponding statistics calculated using data excluded from the k-th fold. We denote by $k(i)$ the number of the fold that containts the i-th observartion.

In the general dynamic setting, as proposed by Jiang and Li (2016) and Thomas and Brunskill (2016), we can construct an AIPW estimator for the welfare function $W(g)$ of a fixed DTR g as follows:[2]

$$\widehat{W}^{AIPW}(g) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{\psi}_{it}^{-k(i)}\left(\underline{g}_t\right)\gamma_t Y_{it} - \sum_{t=1}^{T}\left(\hat{\psi}_{it}^{-k(i)}\left(\underline{g}_t\right) - \hat{\psi}_{i,t-1}^{-k(i)}\left(\underline{g}_{t-1}\right)\right)\cdot \widehat{Q}_t^{\underline{g}_{(t+1):T},-k(i)}\left(H_{it}, D_{it}\right)\right), \tag{A.5}$$

where $\hat{\psi}_{it}^{-k(i)}(\underline{g}_t) \equiv \left(\prod_{s=1}^{t} 1\left\{D_{is} = g_s(H_{is})\right\}\right) / \left(\prod_{s=1}^{t}\hat{e}_t^{-k(i)}\left(H_{is}, g_s\right)\right)$ is an estimator of the sequential propensity weights, and $\widehat{Q}_t^{\underline{g}_{(t+1):T},-k(i)}(h_t, d_t)$ is an estimator of the action-value function (Q-function) for $\underline{g}_{(t+1):T}$:

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

We denote $\hat{\psi}_{i,0}^{-k(i)}(\underline{g}_0) = 1$ when $t = 1$ and $\widehat{Q}_T^{\underline{g}_{(T+1):T},-k(i)}(\cdot,\cdot) = \widehat{Q}_T^{-k(i)}(\cdot,\cdot)$ when $t = T$. The Q-functions $\left\{Q_t^{\underline{g}_{(t+1):T}}(h_t, d_t)\right\}_{t=1,\ldots,T}$ can be estimated by a sequential step-wise algorithm such as the fitted Q-evaluation (Munos and Szepesvári, 2008; Le et al., 2019) from the reinforcement learning literature. The estimator (A.5) generalizes the AIPW of Robins et al. (1994) beyond the static case, and it is a consistent estimator of the population welfare $W(g)$ if either the the propensity weights $\{\psi_t(\cdot)\}_{t=1}^{T}$ or the Q-functions $\{Q_t^{\underline{g}_{t:T}}(\cdot)\}_{t=1}^{T}$ are consistently estimated.

Using the AIPW estimator (A.5), we can estimate the optimal DTR as a solution of the estimated welfare maximization: $\hat{g}^{AIPW} \in \arg\max_{g\in\mathcal{G}} \widehat{W}^{AIPW}(g)$. Remark 5.2 in the main text describes the computational challenge of this method unless the class of DTR $\mathcal{G}$ is not small.

[2]Note that Jiang and Li (2016) and Thomas and Brunskill (2016) consider the estimation of the value of a fixed DTR g, but not the estimation of the optimal DTR.

In what follows, we show the statistical property of $\hat{g}^{AIPW}$. Specifically, we will show the convergence rate of the welfare regret $W^*_{\mathcal{G}} - W(\hat{g}^{AIPW})$. Without loss of generality, we suppose that $\gamma_1 = \cdots = \gamma_T = 1$.

Let $\widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(\cdot,\cdot)$ and $\hat{e}_t^{(n)}(\cdot,\cdot)$, respectively, denote the estimators of the Q-function $Q_t^{\underline{g}_{(t+1):T}}(\cdot,\cdot)$ for $\underline{g}_{(t+1):T}$ and the propensity score $e_t(\cdot,\cdot)$ using size n sample randomly drawn from the population P. We denote $\widehat{Q}_T^{\underline{g}_{(T+1):T},(n)}(\cdot,\cdot) = \widehat{Q}_T^{(n)}(\cdot,\cdot)$ when $t = T$. We suppose that $\{\widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(\cdot,\cdot)\}_{t=1,\ldots,T}$ and $\{\hat{e}_t^{(n)}(\cdot,\cdot)\}_{t=1,\ldots,T}$ satisfy the following assumption.

Assumption E.1. *(i) There exists $\tau > 0$ such that the following holds: For all $t = 1,\ldots,T$, $s = 1,\ldots,t$, and $m \in \{0,1\}$,*

$$\sup_{\underline{d}_{s:t}\in\{0,1\}^{t-s+1}} E\left[\sup_{\underline{g}_{(t+1):T}\in\mathcal{G}_{(t+1):T}}\left(\widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(H_t,d_t) - Q_t^{\underline{g}_{(t+1):T}}(H_t,d_t)\right)^2\right] \times E\left[\left(\frac{1}{\prod_{\ell=s}^{t-m}\hat{e}_\ell^{(n)}(H_\ell,d_\ell)} - \frac{1}{\prod_{\ell=s}^{t-m} e_\ell(H_\ell,d_\ell)}\right)^2\right] = \frac{o(1)}{n^\tau}.$$

(ii) There exists $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$ and $t = 1,\ldots,T$,

$$\sup_{d_t\in\{0,1\},\ \underline{g}_{(t+1):T}\in\mathcal{G}_{(t+1):T}} \widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(H_t,d_t) < \infty \quad \text{and} \quad \sup_{d_t\in\{0,1\}} \hat{e}_t^{(n)}(H_t,d_t) > 0.$$

hold a.s.

Assumption E.1 (i) encompasses the property of double robustness; that is, Assumption E.1 (i) is satisfied if either $\widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(\cdot,\cdot)$ is uniformly consistent or $\prod_{s=t}^T \hat{e}_s^{(n)}(\cdot,\cdot)$ is consistent. As we will see later, the $\sqrt{n}$-consistency of the regret to zero can be achieved when Assumption E.1 (i) holds with $\tau = 1$. This condition is not very restrictive. For example, Assumption E.1 (i) is satisfied when

$$\sup_{d_t\in\{0,1\}} E\left[\sup_{\underline{g}_{(t+1):T}\in\underline{g}_{(t+1):T}}\left(\widehat{Q}_t^{\underline{g}_{(t+1):T},(n)}(H_t,d_t) - Q_t^{\underline{g}_{(t+1):T}}(H_t,d_t)\right)^2\right] = \frac{o(1)}{\sqrt{n}} \text{ and}$$

$$\sup_{\underline{d}_{s:t}\in\{0,1\}^{t-s+1}} E\left[\left(\frac{1}{\prod_{\ell=s}^{t}\hat{e}_\ell^{(n)}(H_\ell,d_\ell)} - \frac{1}{\prod_{\ell=s}^{t} e_\ell(H_\ell,d_\ell)}\right)^2\right] = \frac{o(1)}{\sqrt{n}}$$

hold for all $t = 1, \ldots, T$ and $s = 1, \ldots, t$.

The following theorem shows the convergence rate of the welfare regret $W^*_{\mathcal{G}} - W(\hat{g}^{AIPW})$.

Theorem E.1. *Suppose that Assumptions 2.1–2.4 and E.1 hold. Then*

$$W^*_{\mathcal{G}} - W(\hat{g}^{AIPW}) = O_p(n^{-\min\{1/2, \tau/2\}}). \tag{A.6}$$

Proof. See Appendix F.5. □

When Assumption E.1 holds with $\tau = 1$, the doubly robust estimator $\hat{g}^{AIPW}$ achieves the minimax optimal convergence rate $n^{-1/2}$ of welfare regret. This result is comparable with those of Athey and Wager (2021) and Zhou et al. (2023) who study doubly robust policy learning in the static setting.

E.2 Doubly Robust Estimation with Exogenous Variables

We say that time-varying variables are exogenous when they are not influenced by past treatment assignments. When treatment choice at each stage t depends solely on exogenous variables and past treatment information, we can construct a doubly robust approach to estimate the optimal DTRs with less computational cost. This scenario is prevalent in various contexts. For example, in the context of sequential job training, variables representing exogenous economic conditions (e.g., the unemployment rate in a country where an individual resides) are not influenced by one's job training and are thus considered exogenous. The following assumption formalizes the exogeneity of the covariates.

Assumption E.2. *For any $t = 2, \ldots, T$, $X_t(\underline{d}_{t-1}) = X_t(\underline{d}'_{t-1})$ a.s. for any $\underline{d}_{t-1}, \underline{d}'_{t-1} \in \{0,1\}^{t-1}$.*

Given our focus on using exogenous variables and past treatments exclusively for treatment choice, we redefine the observed and potential history as $H_t \equiv (\underline{D}_{t-1}, \underline{X}_t)$ and $H_t(\underline{d}_{t-1}) \equiv (\underline{d}_{t-1}, \underline{X}_t(\underline{d}_{t-1}))$, respectively, where the history used in treatment choice does not include past outcomes. We suppose that all underlying assumptions (Assumptions 2.1–2.4) for the simultaneous maximization approach hold with this revised definition of H_t.

We denote the conditional expectation function of the weighted outcome $\gamma_t Y_t$ given $\underline{d}_t$ and $\underline{x}_t$ by $\mu_t(\underline{d}_t, \underline{x}_t) \equiv E_P\left[\gamma_t Y_t \mid \underline{D}_t = \underline{d}_t, \underline{X}_t = \underline{x}_t\right]$. Using this function, we can identify $W_t(\underline{g}_t)$ as follows.

Lemma E.2. *Under Assumptions 2.1 and E.2,*

$$W_t\left(\underline{g}_t\right) = \sum_{\underline{d}_t \in \{0,1\}^t} E\left[\mu_t(\underline{d}_t, \underline{X}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\}\right].$$

Proof. See Appendix F.5. □

For each cross-fitting fold k, we estimate $\mu_t(\underline{d}_t, \underline{x}_t)$ and $e_t(h_t, \underline{a}_t)$ by $\hat{\mu}_t^{-k}(\underline{a}_t, s_t)$ and $\hat{e}_t^{-k}(d_t, h_t)$, respectively, using the observations not included in the k-th fold. Any estimation methods, including semi/nonparametric estimators and machine learning methods, can be applied to estimate the nuisance functions $\mu_t(\underline{d}_t, \underline{x}_t)$ and $e_t(h_t, \underline{a}_t)$.

For a fixed DTR g, we construct an AIPW estimator of the welfare function $W(g)$ as

$$\widehat{W}^{DR}(g) \equiv \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{\gamma_t Y_{it} - \hat{\mu}_t^{-k(i)}(\underline{D}_{it}, \underline{X}_{it})}{\prod_{s=1}^{t}\hat{e}_s^{-k(i)}(D_{is}, H_{is})} \cdot \prod_{s=1}^{t} 1\left\{D_{is} = g_s(H_{is})\right\} + \sum_{\underline{d}_t \in \{0,1\}^t}\left\{\hat{\mu}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it}) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_{is})\}\right\}\right].$$

We estimate the optimal DTR by maximizing $\widehat{W}^{DR}(g)$ simultaneously over $\mathcal{G}$. We obtain the DTR estimator $\hat{g}^{DR} \equiv (\hat{g}_1^{DR}, \ldots, \hat{g}_T^{DR})$ as the solution

$$\hat{g}^{DR} \in \arg\max_{g \in \mathcal{G}} \widehat{W}^{DR}(g).$$

In what follows, we show the statistical property of this approach. Let $\hat{\mu}_t^{(n)}(\cdot, \cdot)$ and $\hat{e}_t^{(n)}(\cdot, \cdot)$ denote estimators of the nuisance functions $\mu_t(\cdot, \cdot)$ and $e_t(\cdot, \cdot)$, respectively, using size n sample randomly drawn from the underlying population. We suppose that the following assumption holds.

Assumption E.3. *(i) There exists $\tau > 0$ such that the estimators* $\left\{\hat{\mu}_t^{(n)}(\underline{d}_t, \underline{x}_t) : t = 1, \ldots, T\right\}$

and $\left\{\hat{e}_t^{(n)}(d_t, h_t) : t = 1, \ldots, T\right\}$ *satisfy*

$$\sup_{\underline{d}_t \in \{0,1\}^t} E_{P^n}\left[\left(\hat{\mu}_t^{(n)}(\underline{d}_t, \underline{x}_t) - \mu_t(\underline{d}_t, \underline{x}_t)\right)^2\right] \times E_{P^n}\left[\left(\frac{1}{\prod_{s=1}^t \hat{e}_s^{(n)}(d_s, H_{is})} - \frac{1}{\prod_{s=1}^t e_s(d_s, H_{is})}\right)^2\right] = \frac{o(1)}{n^{\tau'}}$$

for all $t = 1, \ldots, T$.

(ii) There exists $n_0 \in \mathbb{N}$ *such that for any* $n \geq n_0$, $\hat{\mu}_t^{(n)}(\underline{D}_t, \underline{X}_t) < \infty$ *and* $\hat{e}_t^{(n)}(D_t, H_t) > 0$ *hold a.s. for any* t.

As we will see later, the optimal $1/\sqrt{n}$ rate of convergence for the welfare regret of $\hat{g}^{DR}$ can be achieved when Assumption E.3 holds with $\tau' = 1$. This is not a strong or restrictive condition. For example, Assumption E.3 is satisfied when

$$\sup_{\underline{d}_t \in \{0,1\}^t} E_{P^n}\left[\left(\hat{\mu}_t^{(n)}(\underline{d}_t, \underline{X}_{it}) - \mu_t(\underline{d}_t, \underline{X}_{it})\right)^2\right] = \frac{o(1)}{\sqrt{n}} \text{ and}$$

$$\sup_{\underline{d}_t \in \{0,1\}^t} E_{P^n}\left[\left(\frac{1}{\prod_{s=1}^t \hat{e}_s^{(n)}(d_s, H_{is})} - \frac{1}{\prod_{s=1}^t e_s(d_s, H_{is})}\right)^2\right] = \frac{o(1)}{\sqrt{n}}$$

hold for all $t = 1, \ldots, T$. These conditions on the convergence rate of the mean squared errors can be satisfied even with nonparametric estimators with relatively mild conditions (see, e.g., Chernozhukov et al., 2018). Note also that Assumption E.3 encompasses double-robustness of the estimation of the nuisance components.

The following theorem shows the convergence rate of the welfare regret of the doubly robust estimation of the optimal DTR.

Theorem E.3. *Suppose that Assumptions 2.1–2.4, E.2, and E.3 hold with the redefinitions* $H_t \equiv (\underline{D}_{t-1}, \underline{X}_t)$ *and* $H_t(\underline{d}_{t-1}) \equiv (\underline{d}_{t-1}, \underline{X}_t(\underline{d}_{t-1}))$. *Then*

$$W_{\mathcal{G}}^* - W(\hat{g}^{DR}) = O_p(n^{-\min\{1/2, \tau'/2\}}). \tag{A.7}$$

Proof. See Appendix F.5. □

When Assumption E.3 holds with $\tau' = 1$, the doubly robust estimator $\hat{g}^{DR}$ achieves

the minimax optimal convergence rate $1/\sqrt{n}$ of welfare regret. This result is comparable with those of Athey and Wager (2021) and Zhou et al. (2023) who study doubly robust policy learning in static settings.

F Proofs

This appendix provides the proofs of Lemma A.1 and Theorems 3.7, 5.1, C.1, E.1, and E.3.

F.1 Proof of Lemma A.1

We first present the definitions of the VC-dimension of a class of indicator functions and relevant concepts as follows.

Definition F.1 (VC-dimension of a Class of Indicator Functions)**.** *Let $\mathcal{Z}$ be an arbitrary space and $\mathcal{F}$ be a class of indicator functions from $\mathcal{Z}$ to $\{0,1\}$. For a finite sample $S = (z_1, \ldots, z_m)$ of $m \geq 1$ points in $\mathcal{Z}$, we define the set of dichotomies as $\Pi_{\mathcal{F}}(S) \equiv \{(f(z_1), \ldots, f(z_m)) : f \in \mathcal{F}\}$, which is all possible assignments of S by functions in $\mathcal{F}$. We say that S is shattered by $\mathcal{F}$ when $|\Pi_{\mathcal{F}}(S)| = 2^m$; that is, $\mathcal{F}$ realizes all possible dichotomies of S. Then the VC-dimension of $\mathcal{F}$, denoted by $VC(\mathcal{F})$, is defined to be the size of the largest sample S shattered by $\mathcal{F}$, i.e.,*

$$VC(\mathcal{F}) \equiv \max\left\{m : \max_{S=(z_1,\ldots,z_m)\subseteq\mathcal{Z}} |\Pi_{\mathcal{F}}(S)| = 2^m\right\}.$$

We say that $\mathcal{F}$ is a VC-class of indicator functions if $VC(\mathcal{F}) < \infty$.[3]

We next introduce the VC-dimension for a class of subsets. Let $\mathcal{Z}$ be any space, and let $\mathbf{z}_\ell = (z_1, \ldots, z_\ell)$ be a finite set of $\ell \geq 1$ points in $\mathcal{Z}$. Given a class of subsets $\mathcal{G} \subseteq 2^{\mathcal{Z}}$ and a subset $\tilde{\mathbf{z}}$ of $\mathbf{z}_\ell$, we say that $\mathcal{G}$ picks out $\tilde{\mathbf{z}}$ when $\tilde{\mathbf{z}} \cap G = \tilde{\mathbf{z}}$ holds for some $G \in \mathcal{G}$. We say that $\mathcal{G}$ shatters $\mathbf{z}_\ell$ when $|\{\mathbf{z}_\ell \cap G : G \in \mathcal{G}\}| = 2^\ell$ holds, that is all subsets of $\mathbf{z}_\ell$

[3]For example, the class of linear treatment rules

$$\mathcal{G}_t = \left\{1\left\{\beta_{1t}'\mathrm{x}_t + \beta_{2t}'\underline{d}_{t-1} + \beta_{3t}'\underline{y}_{t-1} \geq c_t\right\} : (\beta_{1t}', \beta_{2t}', \beta_{3t}', c_t)' \in \mathbb{R}^{k+2t-1}\right\}.$$

has VC-dimension of at most $k + 2t - 1$.

are picked out by $\mathcal{G}$. The VC-dimension of the class of subsets $\mathcal{G}$, denoted by $VC(\mathcal{G})$, is defined as the cardinality of the largest subset $\mathbf{z}_\ell$ contained in $\mathcal{Z}$ and shattered by $\mathcal{G}$, i.e.,

$$VC(\mathcal{G}) \equiv \max\{\ell : \max_{\mathbf{z}_\ell \subseteq \mathcal{Z}} |\{\mathbf{z}_\ell \cap G : G \in \mathcal{G}\}| = 2^\ell\}.$$

We say that a class of subsets $\mathcal{G}$ is a VC-class of subsets if $VC(\mathcal{G}) < \infty$.

We next introduce a concept of the subgraph of a real-valued function $f : \mathcal{Z} \to \mathbb{R}$ that is the set

$$\mathrm{SG}(f) \equiv \{(z,t) \in \mathcal{Z} \times \mathbb{R} : t \leq f(z)\}.$$

Let $\mathrm{SG}(\mathcal{F}) \equiv \{\mathrm{SG}(f) : f \in \mathcal{F}\}$ be a collection of subgraphs over a class of functions $\mathcal{F}$. We here consider the VC-dimension of $\mathrm{SG}(\mathcal{F})$ as a complexity measure of $\mathcal{F}$. Note that in the case of $\mathcal{F}$ being a class of indicator functions, the VC-dimension of $\mathrm{SG}(\mathcal{F})$ corresponds to the VC-dimension of $\mathcal{F}$ in the sense of Definition F.1. We say that a class of functions $\mathcal{F}$ is a VC-subgraph class of functions if $VC(\mathrm{SG}(\mathcal{F})) < \infty$.

The following lemmas are auxiliary lemmas for Lemma A.1.

Lemma F.2. *(Sauer's lemma; see, for example, Theorem 3.6.2 of Giné and Nickl (2016)) Let $\mathcal{Z}$ be any space, and $(z_1, \ldots, z_\ell)$ be a finite set of $\ell \geq 1$ points in $\mathcal{Z}$. Let $\mathcal{G}$ be a VC-class of subsets in $\mathcal{Z}$ with $VC(\mathcal{G}) = v < \infty$. Let $\Delta_\ell(\mathcal{G}, (z_1, \ldots, z_\ell))$ denote the number of subsets of $(z_1, \ldots, z_\ell)$ that are picked out by $\mathcal{G}$, i.e.,*

$$\Delta_\ell(\mathcal{G}, (z_1, \ldots, z_\ell)) \equiv |\{(z_1, \ldots, z_\ell) \cap G : G \in \mathcal{G}\}|.$$

Then the following holds:

$$\max_{(z_1,\ldots,z_\ell)\subseteq\mathcal{Z}} \Delta_\ell(\mathcal{G}, (z_1, \ldots, z_\ell)) \leq \sum_{j=0}^{v} \binom{\ell}{j} \leq \left(\frac{\ell e}{v}\right)^v.$$

Lemma F.3. *Let $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2$ be any product space, and $\mathcal{G}$ be a class of indicator functions from $\mathcal{Z}_2$ to $\{0,1\}$. Suppose that $\mathcal{G}$ has VC-dimension $v \geq 0$ in the sense of*

Definition F.1. Fix a function f on $\mathcal{Z}$, and define a class of functions on $\mathcal{Z}$:

$$\mathcal{F}_{\mathcal{G}} = \{f \cdot g : g \in \mathcal{G}\}.$$

Then $\mathcal{F}_{\mathcal{G}}$ is a VC-subgraph class of functions with $VC(SG(\mathcal{F}_{\mathcal{G}})) \leq v$.

Proof. We prove the statement by contradiction. Suppose that there exist some $(v+1)$-points $\{(z_1, t_1), \ldots, (z_{v+1}, t_{v+1})\} \equiv \{(z_{1,1}, z_{2,1}, t_1), \ldots, (z_{1,v+1}, z_{2,v+1}, t_{v+1})\} \subset \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathbb{R}$ that are shattered by $\mathrm{SG}(\mathcal{F}_{\mathcal{G}})$.

When $t \leq f(z) \wedge 0$ or $t > f(z) \vee 0$ for some $(z, t) \in \{(z_1, t_1), \ldots, (z_{v+1}, t_{v+1})\}$, $\mathrm{SG}(\mathcal{F}_{\mathcal{G}})$ cannot pick out $\{(z_1, t_1), \ldots (z_{v+1}, t_{v+1})\} \setminus \{(z, t)\}$ or $\{(z, t)\}$. Thus, we need to consider only the case that $(f(z) \wedge 0) < t \leq (f(z) \vee 0)$ for all $(z, t) \in \{(z_1, t_1), \ldots (z_{v+1}, t_{v+1})\}$. In the remaining case, we indicate $\delta_j = 1$ if $t_j \leq f(z_j)$ and $\delta_j = 0$ otherwise. Since the VC-dimension of $\mathcal{G}$ is at most v in the sense of Definition F.1, there exists a subset $S \equiv (\tilde{z}_{2,1}, \ldots, \tilde{z}_{2,m})$ (for some $m > 0$) of $\{z_{2,1}, \ldots, z_{2,v+1}\}$ such that $(g(\tilde{z}_{2,1}), \ldots, g(\tilde{z}_{2,m})) \neq (1, \ldots, 1)$ and $(g(z_{2,1}), \ldots, g(z_{2,v+1}) \backslash (g(\tilde{z}_{2,1}), \ldots, g(\tilde{z}_{2,m})) \neq (0, \ldots, 0)$ for any $g \in \mathcal{G}$. Then $\mathrm{SG}(\mathcal{F}_{\mathcal{G}})$ cannot pick out the following subset:

$$\{(z_j, t_j) : (z_{2,j} \in S \text{ and } \delta_j = 1) \text{ or } (z_{2,j} \notin S \text{ and } \delta_j = 0)\},$$

because this set of points could be contained in $\mathrm{SG}(f \cdot g)$ only when $\mathrm{sign}(t_j) = \mathrm{sign}(g(z_{2,j}) - 1/2)$ for all $j = 1, \ldots, v+1$. This contradicts the assumption that $\{(z_1, t_1), \ldots (z_{v+1}, t_{v+1})\} \subset \mathcal{Z} \times \mathbb{R}$ is shattered by $\mathrm{SG}(\mathcal{F}_{\mathcal{G}})$. □

Proof of Lemma A.1. We prove for the case that $s = 1$ and $t = T$. The result follows for the remaining cases by a similar argument. Let m be an arbitrary integer and $(z_1, \ldots, z_m)$ be m arbitrary points on $\mathcal{Z}$. For each t, fixing $g_s \in \mathcal{G}_s$ for all $s \neq t$, define a class of functions

$$\widetilde{\mathcal{F}}_t \equiv \{f(z) = 1\{g_1(h_1) = d_1, \ldots, g_T(h_T) = d_T\} \cdot r(z) : g_t \in \mathcal{G}_t\},$$

and, fixing $g_s \in \mathcal{G}_s$ for all $s > t$, define

$$\widetilde{\mathcal{F}}_{1:t} \equiv \{f(z) = 1\{g_1(h_1) = d_1, \ldots, g_T(h_T) = d_T\} \cdot r(z) : (g_1, \ldots, g_t) \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t\}.$$

We first consider $\widetilde{\mathcal{F}}_1$, or equivalently $\widetilde{\mathcal{F}}_{1:1}$. Applying Lemma F.3 to $\widetilde{\mathcal{F}}_1$ shows that $\widetilde{\mathcal{F}}_1$ is a VC-subgraph of functions with $VC(\mathrm{SG}(\widetilde{\mathcal{F}}_1)) \leq v_1$. Therefore, from Lemma F.2, $\mathrm{SG}(\widetilde{\mathcal{F}_1})$ can pick out at most $O(m^{v_1})$ subsets from $(z_1, \ldots, z_m)$.

Next we study $\widetilde{\mathcal{F}}_2$ and then $\widetilde{\mathcal{F}}_{1:2}$. Let $(z_1, \ldots, z_{m'})$ be an arbitrary subset picked out by $\mathrm{SG}(\tilde{f}_1)$ where $\tilde{f}_1 \in \widetilde{\mathcal{F}}_{1:1}$ has a fixed $g_1 \in \mathcal{G}_1$. Lemmas F.2 and F.3 show that $\mathrm{SG}(\widetilde{\mathcal{F}}_2)$ can pick out at most $O(m^{v_2})$ subsets from $(z_1, \ldots, z_{m'})$. Because $\mathrm{SG}(\widetilde{\mathcal{F}}_2)$ can pick out at most $O(m^{v_2})$ subsets from each subset of $(z_1, \ldots, z_m)$ and $\mathrm{SG}(\widetilde{\mathcal{F}}_{1:1})$ can pick out at most $O(m^{v_1})$ subsets from $(z_1, \ldots, z_m)$, by varying (g_1, g_2) over $\mathcal{G}_1 \times \mathcal{G}_2$, $\mathrm{SG}(\widetilde{\mathcal{F}}_{1:2})$ picks out at most $O(m^{v_1+v_2})$ subsets from $(z_1, \ldots, z_m)$.

For $s \geq 2$, suppose that $\widetilde{\mathcal{F}}_{1:s-1}$ can pick out at most $O(m^{\sum_{t=1}^{s-1} v_t})$ subsets from $(z_1, \ldots, z_m)$. Let $(z_1, \ldots, z_{m'})$ be an arbitrary subset picked out by $\mathrm{SG}(\tilde{f}_{s-1})$ where $\tilde{f}_{s-1} \in \widetilde{\mathcal{F}}_{1:s-1}$ has fixed $g_1, \ldots, g_{s-1}$. From $(z_1, \ldots, z_{m'})$, $\mathrm{SG}(\tilde{\mathcal{F}}_s)$ can pick out at most $O(m^{v_s})$ subsets. Combining this result with the fact that $\widetilde{\mathcal{F}}_{1:s-1}$ can pick out at most $O(m^{\sum_{t=1}^{s-1} v_t})$ subsets from $(z_1, \ldots, z_m)$ leads to the conclusion that $\widetilde{\mathcal{F}}_{1:s}$ picks out at most $O(m^{\sum_{t=1}^{s} v_t})$ subsets from $(z_1, \ldots, z_m)$.

Recursively, we can prove that $\widetilde{\mathcal{F}}_{1:T}$ picks out at most $O(m^{\sum_{t=1}^{T} v_t})$ subsets from $(z_1, \ldots, z_m)$. Hence, $\mathrm{SG}(\widetilde{\mathcal{F}}_{1:T})$ is a VC-subgraph class of functions with VC-dimension less than or equal to $\sum_{t=1}^{T} v_t$. □

F.2 Proof of Theorem 3.7.

This appendix present the proof of Theorem 3.7. The following is its auxiliary lemma, where we use the same strategy as the proofs of Theorem 2 of Massart et al. (2006) and Theorem 2.2 of Kitagawa and Tetenov (2018), but extend it to the dynamic treatment setting.

Lemma F.4. *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* $\mathcal{G}$*. Fix* $t \in \{1, \ldots, T\}$*, and let* $\gamma_t = 1$ *and* $\gamma_s = 0$

for all $s \neq t$. Then, for any DTR $\hat{g} \in \mathcal{G}$ as a function of $(Z_1, \ldots, Z_n)$,

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

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

Proof. The proof follows by constructing a specific subclass of $\mathcal{P}\left(M, \kappa, \mathcal{G}\right)$, for which the worst-case average welfare regret can be bounded from below. We here prove the statement for the lemma in the case that $t = T$ (i.e., $\gamma_T = 1$ and $\gamma_s = 0$ for $s \neq T$). The proof follows for the remaining cases by a similar argument. For simplicity, we normalize the support of the potential outcomes to $Y_t\left(\underline{d}_t\right) \in [-1/2, 1/2]$ for all $\underline{d}_t \in \{0,1\}^t$ and $t = 1, \ldots, T$. We also suppose that $X_t(\underline{d}_{t-1}) = X_t(\underline{d}'_{t-1})$ for any $t \geq 2$ and any $\underline{d}_{t-1}, \underline{d}'_{t-1} \in \{0,1\}^{t-1}$; that is, the covariates do not depend on the past treatments. Let $\mathbf{1}_T$ denote a T-dimensional vector of ones.

We construct a specific subclass $\mathcal{P}^* \subset \mathcal{P}(\mathbf{1}_T, \kappa)$ as follows. Let $\tilde{Z} \equiv ((D_t, X_t, Y_t)_{t=1}^{T-1}, D_T, X_T)$, which is a vector of all the observed variables excluding Y_T, and denote its space by $\widetilde{\mathcal{Z}}$. Let $\tilde{z}_1, \ldots, \tilde{z}_{v_{1:T}}$ be $v_{1:T}$ points in $\widetilde{\mathcal{Z}}$ such that a set $\{(\tilde{z}_1, 1/2), \ldots, (\tilde{z}_{v_{1:T}}, 1/2)\}$ is shattered by a collection of indicator functions

$$\left\{f\left(z\right) = 1\left\{g_1\left(h_1\right) = d_1, \cdots, g_T\left(h_T\right) = d_T\right\} : \left(g_1, \ldots, g_T\right) \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_T\right\}$$

in the sense of Definition F.1. For $j = 1, \ldots, v_{1:T}$, denote $\tilde{z}_j = \left((d_{tj}, x_{tj}, y_{tj})_{t=1}^{T-1}, d_{Tj}, x_{Tj}\right) \in \widetilde{\mathcal{Z}}$. For any $P \in \mathcal{P}^*$, we suppose for the marginal distributions of $\tilde{Z}$ on $\widetilde{\mathcal{Z}}$ that $P\left(\tilde{Z} = \tilde{z}_j\right) = 1/v_{1:T}$ for each $j = 1, \ldots, v_{1:T}$. Let $\mathbf{b} = (b_1, \ldots, b_{v_{1:T}}) \in \{0,1\}^{v_{1:T}}$ be a bit vector that indexes a member of $\mathcal{P}^*$. Hence $\mathcal{P}^*$ consists of $2^{v_{1:T}}$ distinct DGPs. For each $j = 1, \ldots, v_{1:T}$, depending on $\mathbf{b}$, we construct the following conditional distribution of $Y_T\left(\underline{d}_T\right)$ given $\tilde{Z} = \tilde{z}_j$: if $b_j = 1$,

$$Y_T\left(\underline{d}_{Tj}\right) = \begin{cases} 1/2 & \text{w.p. } 1/2 + \delta \\ -1/2 & \text{w.p. } 1/2 - \delta \end{cases},$$

otherwise

$$Y_T\left(\underline{d}_{Tj}\right) = \begin{cases} 1/2 & \text{w.p. } 1/2-\delta \\ -1/2 & \text{w.p. } 1/2+\delta \end{cases},$$

where $\underline{d}_{Tj}$ is the history of the realized treatments from stage 1 to T when $\tilde{Z} = \tilde{z}_j$, and $\delta \in [0, 1/2]$ is chosen properly in a later step of the proof. When $b_j = 1$, $E_P\left[Y_T\left(\underline{d}_{Tj}\right) \mid \tilde{Z} = \tilde{z}_j\right] = \delta$; otherwise, $E_P\left[Y_T\left(\underline{d}_{Tj}\right) \mid \tilde{Z} = \tilde{z}_j\right] = -\delta$. For conditional distributions of the other potential outcomes $Y_T\left(\underline{d}_T\right)$ given $\tilde{Z} = \tilde{z}_j$, we set $Y_T\left(\underline{d}_T\right) = 0$ with probability 1 if $\underline{d}_T \neq \underline{d}_{Tj}$.

When $\mathbf{b}$ is known, an optimal DTR, denoted by $g^*_{\mathbf{b}} = \left(g^*_{1,\mathbf{b}}, \ldots, g^*_{T,\mathbf{b}}\right)$, is such that

$$\left(g^*_{1,\mathbf{b}}(h_{1j}), \ldots, g^*_{T,\mathbf{b}}(h_{Tj})\right) = \begin{cases} \underline{d}_{Tj} & \text{if } b_j = 1 \\ (1-d_{1j}, \cdots, 1-d_{Tj}) & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, v_{1:T}$, where h_{tj} is the history information in $\tilde{z}_j$ up to stage t. Such a DTR is feasible in $\mathcal{G}$. Then, the optimized social welfare given $\mathbf{b}$ is

$$W\left(g^*_{\mathbf{b}}\right) = \frac{1}{v_{1:T}} \delta \sum_{j=1}^{v_{1:T}} b_j.$$

Let $\hat{g} = (\hat{g}_1, \ldots, \hat{g}_T) : \mathcal{H}_1 \times \cdots \times \mathcal{H}_T \mapsto \{0,1\}^T$ be an arbitrary DTR depending on the sample $(Z_1, \ldots, Z_n)$, and let $\hat{\mathbf{b}} \in \{0,1\}^{v_{1:T}}$ be a binary vector such that its j-th element is given by

$$\hat{b}_j = 1\left\{\hat{g}_1\left(h_{1j}\right) = d_{1j}, \ldots, \hat{g}_T\left(h_{Tj}\right) = d_{Tj}\right\}.$$

We define by $g(\mathbf{b})$ a prior of $\mathbf{b}$ such that $b_1, \ldots, b_{v_{1:T}}$ are i.i.d and $b_1 \sim \text{Ber}(1/2)$.

Then the maximum average welfare regret on $\mathcal{P}(\mathbf{1}_T, \kappa)$ satisfies the following:

$$\begin{aligned} &\sup_{P \in \mathcal{P}(\mathbf{1}_T, \kappa)} E_{P^n}\left[W^*_{\mathcal{G}} - W\left(\hat{g}\right)\right] \\ &\geq \sup_{P_{\mathbf{b}} \in \mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W\left(g^*_{\mathbf{b}}\right) - W\left(\hat{g}\right)\right] \geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W\left(g^*_{\mathbf{b}}\right) - W\left(\hat{g}\right)\right] dg(\mathbf{b}) \\ &\geq \delta \int_{\mathbf{b}} \int_{Z_1, \ldots, Z_n} P_{\tilde{Z}}\left(\left\{b(\tilde{Z}) \neq \hat{b}(\tilde{Z})\right\}\right) dP^n_{\mathbf{b}}\left(Z_1, \ldots, Z_n\right) dg(\mathbf{b}) \end{aligned}$$

$$\geq \inf_{\hat{g}\in\mathcal{G}} \delta \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_{\tilde{Z}} \left(\left\{ b(\tilde{Z}) \neq \hat{b}(\tilde{Z}) \right\} \right) dP_{\mathbf{b}}^{n} \left(Z_1, \ldots, Z_n \right) dg(\mathbf{b}),$$

where $P_{\tilde{Z}}$ is a probability measure of $\tilde{Z}$, and $b(\tilde{Z})$ and $\hat{b}(\tilde{Z})$ are elements of $\mathbf{b}$ and $\hat{\mathbf{b}}$ such that $b\left(\tilde{z}_j\right) = b_j$ and $\hat{b}\left(\tilde{z}_j\right) = \hat{b}_j$, respectively. Note that the above minimization problem can be seen as the minimization of the Bayes risk when the loss function corresponds to the classification error for predicting the binary random variable $b(\tilde{Z})$. Hence, the risk is minimized by the Bayes classifier such that for each $j = 1, \ldots, J$,

$$\hat{b}^{*}\left(\tilde{z}_j\right) = \begin{cases} 1 & \text{if } g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right) \geq 1/2 \\ 0 & \text{otherwise} \end{cases},$$

where $g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right)$ is the posterior distribution for $b_j = 1$. This Bayes classifier is achieved by a DTR $\hat{g}^{*} \equiv \left(\hat{g}_1^{*}, \ldots, \hat{g}_T^{*}\right)$ that satisfies for $j = 1, \ldots, j$,

$$\left(\hat{g}_1^{*}\left(h_{1j}\right), \ldots, \hat{g}_T^{*}\left(h_{Tj}\right)\right) = \begin{cases} \underline{d}_{Tj} & \text{if } g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right) \geq 1/2 \\ \left(1 - d_{1j}, \ldots, 1 - d_{Tj}\right) & \text{otherwise} \end{cases}.$$

Note that $\hat{g}^{*}$ is feasible in $\mathcal{G}$.

Then, using $\hat{g}^{*}$, the minimized risk is given by

$$\begin{aligned} &\delta \int_{Z_1,\ldots,Z_n} E_{\tilde{Z}} \left[\min \left\{ g\left(b\left(\tilde{Z}\right) = 1 \mid Z_1, \ldots, Z_n\right), 1 - g\left(b\left(\tilde{Z}\right) = 1 \mid Z_1, \ldots, Z_n\right) \right\} \right] d\tilde{P}^{n} \\ =&\frac{1}{v_{1:T}} \delta \int_{Z_1,\ldots,Z_n} \sum_{j=1}^{v_{1:T}} \left[\min \left\{ g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right), 1 - g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right) \right\} \right] d\tilde{P}^{n}, \quad \text{(A.8)} \end{aligned}$$

where $\tilde{P}$ is the marginal likelihood of $\left\{ \left\{ Y_{it}\left(\underline{d}_t\right) \right\}_{\underline{d}_t \in \{0,1\}^t}, \{D_{it}, X_{it}\} \right\}_{i=1,\ldots,n; t=1,\ldots,T}$ with prior $g(\mathbf{b})$.

For each $j = 1, \ldots, v_{1:T}$, let

$$k_j^{+} = \# \left\{ i : \tilde{Z}_i = \tilde{z}_j, Y_{iT} = \frac{1}{2} \right\},$$
$$k_j^{-} = \# \left\{ i : \tilde{Z}_i = \tilde{z}_j, Y_{iT} = -\frac{1}{2} \right\}.$$

Then the posteriors for $b_j = 1$ can be written as

$$g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right) = \begin{cases} \frac{1}{2} & \text{if } k_j^+ + k_j^- = 0 \\ \frac{\left(\frac{1}{2}+\delta\right)^{k_j^+}\left(\frac{1}{2}-\delta\right)^{k_j^-}}{\left(\frac{1}{2}+\delta\right)^{k_j^+}\left(\frac{1}{2}-\delta\right)^{k_j^-}+\left(\frac{1}{2}+\delta\right)^{k_j^-}\left(\frac{1}{2}-\delta\right)^{k_j^+}} & \text{otherwise.} \end{cases}$$

Hence, the following holds:

$$\begin{aligned} & \min\left\{g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right), 1 - g\left(b_j = 1 \mid Z_1, \ldots, Z_n\right)\right\} \\ = & \frac{\min\left\{\left(\frac{1}{2}+\delta\right)^{k_j^+}\left(\frac{1}{2}-\delta\right)^{k_j^-}, \left(\frac{1}{2}+\delta\right)^{k_j^-}\left(\frac{1}{2}-\delta\right)^{k_j^+}\right\}}{\left(\frac{1}{2}+\delta\right)^{k_j^+}\left(\frac{1}{2}-\delta\right)^{k_j^-}+\left(\frac{1}{2}+\delta\right)^{k_j^-}\left(\frac{1}{2}-\delta\right)^{k_j^+}} \\ = & \frac{\min\left\{1, \left(\frac{\frac{1}{2}+\delta}{\frac{1}{2}-\delta}\right)^{k_j^+-k_j^-}\right\}}{1+\left(\frac{\frac{1}{2}+\delta}{\frac{1}{2}-\delta}\right)^{k_j^+-k_j^-}} \\ = & \frac{1}{1+a^{\left|k_j^+-k_j^-\right|}}, \text{ where } a = \frac{1+2\delta}{1-2\delta} > 1. \end{aligned} \tag{A.9}$$

Since

$$k_j^+ - k_j^- = \sum_{i:\tilde{Z}_i = \tilde{z}_j} 2Y_{Ti},$$

plugging (A.9) into (A.8) yields

$$\begin{aligned} (A.8) = \frac{1}{v_{1:T}}\delta\sum_{j=1}^{v_{1:T}} E_{\tilde{P}^n}\left[\frac{1}{1+a^{\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\right|}}\right] &\geq \frac{\delta}{2v_{1:T}}\sum_{j=1}^{v_{1:T}} E_{\tilde{P}^n}\left[\frac{1}{a^{\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\right|}}\right] \\ &\geq \frac{\delta}{2v_{1:T}}\sum_{j=1}^{v_{1:T}} a^{-E_{\tilde{P}^n}\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\right|}, \end{aligned}$$

where $E_{\tilde{P}^n}[\cdot]$ is the expectation with respect to the marginal likelihood of

$$\left\{\left\{Y_{it}\left(\underline{d}_t\right)\right\}_{\underline{d}_t \in \{0,1\}^t}, D_{it}, X_{it}\right\}_{i=1,\ldots,n;t=1,\ldots,T}.$$

The first inequality follows by $a > 1$, and the second inequality follows by Jensen's inequality. Given our prior distribution for $\mathbf{b}$, for each $\underline{d}_T \in \{0,1\}^T$, the marginal distribution of $Y_{iT}\left(\underline{d}_T\right)$ is $P\left(Y_{iT}\left(\underline{d}_T\right) = 1/2\right) = P\left(Y_{iT}\left(\underline{d}_T\right) = -1/2\right) = 1/2$ if there exist $\underline{d}_{Tj}$ among

$\underline{d}_{T1}, \ldots, \underline{d}_{Tv_{1:T}}$ such that $\underline{d}_{Tj} = \underline{d}_T$; otherwise, $P\left(Y_{iT}\left(\underline{d}_T\right) = 0\right) = 1$. Thus, we have

$$\begin{aligned} E_{\tilde{P}^n}\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\right| &= E_{\tilde{P}^n}\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\left(\underline{d}_{Tj}\right)\right| \\ &= \sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)\left(\frac{1}{v_{1:T}}\right)^k\left(1-\frac{1}{v_{1:T}}\right)^{n-k} E\left|B\left(k,\frac{1}{2}\right)-\frac{k}{2}\right|, \end{aligned}$$

where $B\left(k, 1/2\right)$ is the binomial random variable with parameters k and $1/2$. By the Cauchy-Schwarz inequality, it follows that

$$E\left|B\left(k,\frac{1}{2}\right)-\frac{1}{2}\right| \leq \sqrt{E\left(B\left(k,\frac{1}{2}\right)-\frac{k}{2}\right)^2} = \sqrt{\frac{k}{4}}.$$

Thus, we obtain

$$\begin{aligned} E_{\tilde{P}^n}\left|\sum_{i:\tilde{Z}_i=\tilde{z}_j} 2Y_{iT}\right| &\leq \sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)\left(\frac{1}{v_{1:T}}\right)^k\left(1-\frac{1}{v_{1:T}}\right)^{n-k}\sqrt{\frac{k}{4}} \\ &= E\sqrt{\frac{B\left(n,\frac{1}{v_{1:T}}\right)}{4}} \\ &\leq \sqrt{\frac{n}{4v_{1:T}}}, \end{aligned}$$

where the last inequality follows by Jensen's inequality. Hence, the Bayes risk is bounded from below by

$$\begin{aligned} \frac{\delta}{2}a^{-\sqrt{\frac{n}{4v_{1:T}}}} &\geq \frac{\delta}{2}\exp\left\{-\left(a-1\right)\sqrt{\frac{n}{4v_{1:T}}}\right\} \\ &= \frac{\delta}{2}\exp\left\{-\frac{4\delta}{1-2\delta}\sqrt{\frac{n}{4v_{1:T}}}\right\}, \end{aligned}$$

where the inequality follows from the fact that $1+x \leq e^x$ for any x. This lower bound on the Bayes risk has the slowest convergence rate when δ is set to be proportional to $n^{-1/2}$. Specifically, letting $\delta = \sqrt{v_{1:T}/n}$, we have

$$\frac{\delta}{2}\exp\left\{-\frac{4\delta}{1-2\delta}\sqrt{\frac{n}{4v_{1:T}}}\right\} = \frac{1}{2}\sqrt{\frac{v_{1:T}}{n}}\exp\left\{-\frac{2}{1-2\delta}\right\} \geq \frac{1}{2}\sqrt{\frac{v_{1:T}}{n}}\exp\left(-4\right) \text{ if } 1-2\delta \geq \frac{1}{2}.$$

The condition $1-2\delta \geq 1/2$ is equivalent to $n \geq 16 v_{1:T}$. Multiplying the lower bound by M_T gives

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

for all $n \geq 16 v_{1:T}$.

The proof is valid irrespective of whether Assumption 3.1 holds for a pair $(P, \mathcal{G})$ with any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$ or not. □

Proof of Theorem 3.7. The result immediately follows by setting

$$t = \underset{s \in \{1,\ldots,T\}}{\arg\max} \gamma_s M_s \sqrt{\frac{v_{1:s}}{n}}$$

in the statement of Lemma F.4. □

F.3 Proof of Theorem 5.1

We derive uniform upper bounds on the worst-case average welfare regrets of the two DEWM methods in the case where estimated propensity scores are used instead of true ones.

Proof of Theorem 5.1 (i). Let $P \in \mathcal{P}_e \cap \mathcal{P}(M, \kappa, \mathcal{G})$ be fixed. Define $\hat{W}_{nt}(\underline{g}_t) \equiv n^{-1}\sum_{i=1}^n \hat{w}_t^S(Z_i, \underline{g}_t)$ and $\hat{W}_n(g) \equiv \sum_{t=1}^T \hat{W}_{nt}(\underline{g}_t)$, which are estimators of $W_t(\underline{g}_t)$ and $W(g)$, respectively. It follows for any $g \in \mathcal{G}$ that

$$\begin{aligned}
W(g) - W\left(\hat{g}_e^S\right) &\leq W_n(g) - \hat{W}_n(g) - W_n\left(\hat{g}_e^S\right) + \hat{W}_n\left(\hat{g}_e^S\right) \\
&\quad + W(g) - W\left(\hat{g}_e^S\right) + W_n\left(\hat{g}_e^S\right) - W_n(g) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T \sum_{\underline{d}_t \in \{0,1\}^t} \left[\left(\frac{\gamma_t Y_{it} \cdot 1\left\{\underline{D}_{it} = \underline{d}_t\right\}}{\prod_{s=1}^t e_s\left(d_s, H_{is}\right)} - \frac{\gamma_t Y_{it} \cdot 1\left\{\underline{D}_{it} = \underline{d}_t\right\}}{\prod_{s=1}^t \hat{e}_s\left(d_s, H_{is}\right)}\right)\right. \\
&\quad \left.\times \left(\prod_{s=1}^t 1\left\{g_s\left(H_{is}\right) = d_s\right\} - \prod_{s=1}^t 1\left\{\hat{g}_{e,s}^S\left(H_{is}\right) = d_s\right\}\right)\right] \\
&\quad + W(g) - W_n(g) + W_n\left(\hat{g}_e^S\right) - W\left(\hat{g}_e^S\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \left(\frac{1}{n} \sum_{i=1}^{n} \left| \tau_t \left(\underline{d}_t, H_{it} \right) - \hat{\tau}_t \left(\underline{d}_t, H_{it} \right) \right| \right) + 2 \sup_{g \in \mathcal{G}} \left| W_n \left(g \right) - W \left(g \right) \right| \\
&\leq \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \left(\frac{1}{n} \sum_{i=1}^{n} \left| \tau_t \left(\underline{d}_t, H_{it} \right) - \hat{\tau}_t \left(\underline{d}_t, H_{it} \right) \right| \right) \\
&+ 2 \sum_{t=1}^{T} \sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} \left| W_{nt}(\underline{g}_t) - W_t(\underline{g}_t) \right|.
\end{aligned}
$$

The first inequality follows from the fact that $\hat{g}_e^S$ maximizes $\hat{W}_n(\cdot)$ over $\mathcal{G}$. The second inequality follows from the fact that

$$
\left| \prod_{s=1}^{t} 1 \left\{ g_s \left(H_{is} \right) = d_s \right\} - \prod_{s=1}^{t} 1 \left\{ \hat{g}_{e,s}^S \left(H_{is} \right) = d_s \right\} \right| \leq 1.
$$

Thus, the average welfare regret can be bounded from above by

$$
\begin{aligned}
E_{P^n} \left[W_{\mathcal{G}}^* - W \left(\hat{g}_e^S \right) \right] \leq & \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} E_{P^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left| \tau_t \left(\underline{d}_t, H_{it} \right) - \hat{\tau}_t \left(\underline{d}_t, H_{it} \right) \right| \right] \\
& + 2 \sum_{t=1}^{T} \sup_{\underline{g}_t \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_t} E_{P^n} \left[\left| W_{nt}(\underline{g}_t) - W_t(\underline{g}_t) \right| \right].
\end{aligned}
$$

Therefore, by the same argument as in the proof of Theorem 3.6 (i) and from Assumption 5.1 (i), the average welfare regret is bounded from above as

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

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

Before proceeding to the proof of Theorem 5.1 (ii), we define

$$
\begin{aligned}
\Delta \tilde{Q}_{t,e} &\equiv \tilde{Q}_t \left(g_t^*, \ldots, g_T^* \right) - \tilde{Q}_t \left(\hat{g}_{t,e}^B, \ldots, \hat{g}_{T,e}^B \right), \\
\Delta \tilde{Q}_{t,e}^{\dagger} &\equiv \tilde{Q}_t \left(g_t^*, \hat{g}_{t+1,e}^B, \ldots, \hat{g}_{T,e}^B \right) - \tilde{Q}_t \left(\hat{g}_{t,e}^B, \ldots, \hat{g}_{T,e}^B \right), \\
\check{Q}_{nt} \left(g_t, \ldots, g_T \right) &\equiv E_n \left[\hat{q}_t \left(Z, g_t, \ldots, g_T \right) \right]
\end{aligned}
$$

$$= \sum_{s=t}^{T} E_n \left[\frac{(\prod_{\ell=t}^{s} 1\{g_\ell(H_\ell) = D_\ell\})\gamma_s Y_s}{\prod_{\ell=t}^{s} \hat{e}_\ell(D_\ell, H_\ell)} \right],$$

where $\tilde{Q}$ is defined in (16). The following lemmas will be used in the proof of Theorem 5.1 (ii).

Lemma F.5. *Suppose that Assumptions 2.1, 2.2, and 2.4 hold for any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$, that Assumption 2.3 holds for $\mathcal{G}$, that Assumption 3.1 holds for a pair $(P, \mathcal{G})$ with any $P \in \mathcal{P}(M, \kappa, \mathcal{G})$, and that Assumption 5.1 (ii) holds for any $P \in \mathcal{P}_e$. Then, for any $P \in \mathcal{P}(M, \kappa, \mathcal{G}) \cap \mathcal{P}_e$, the following hold:*

(i) for $t = 1, \ldots, T$,

$$E_{P^n}\left[\Delta\tilde{Q}^{\dagger}_{t,e}\right] \leq C\left(\sum_{s=t}^{T} \frac{\gamma_s M_s}{\prod_{\ell=t}^{s} \kappa_\ell}\right)\sqrt{\frac{\sum_{s=t}^{T} v_s}{n}} + O\left(\xi_n^{-1}\right),$$

where C is the same constant term as introduced in Lemma A.2;

(ii) for $t = 1, \ldots, T-1$ and $s = t+1, \ldots, T$,

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

(iii)

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

Proof. (i) It follows for any $\tilde{g}_t \in \mathcal{G}_t$ that

$$\begin{aligned}
&\tilde{Q}_t\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_t\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) \\
&\leq \tilde{Q}_{nt}\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \check{Q}_{nt}\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_{nt}\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) + \check{Q}_{nt}\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) \\
&+ \tilde{Q}_t\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_t\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) + \tilde{Q}_{nt}\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_{nt}\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{s=t}^{T}\sum_{\underline{d}_{t:s} \in \{0,1\}^{t-s+1}} \left[\left(\frac{1\{\underline{D}_{t:s} = \underline{d}_{t:s}\}\gamma_s Y_s}{\prod_{\ell=t}^{s} e_\ell(d_\ell, H_\ell)} - \frac{1\{\underline{D}_{t:s} = \underline{d}_{t:s}\}\gamma_s Y_s}{\prod_{\ell=t}^{s} \hat{e}_\ell(d_\ell, H_\ell)}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\times \left(1\left\{\tilde{g}_t(H_{it}) = d_t\right\} - 1\left\{\hat{g}^B_{t,e}\left(H_{it}\right) = d_t\right\}\right)\Big] \\
&+ \tilde{Q}_t\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_t\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) + \tilde{Q}_{nt}\left(\tilde{g}_t, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_{nt}\left(\hat{g}^B_{t,e}, \ldots, \hat{g}^B_{T,e}\right) \\
&\leq \sum_{\underline{d}_{t:T} \in \{0,1\}^{T-t+1}} \left(\frac{1}{n} \sum_{i=1}^{n} \left| \hat{\eta}_t^{-k}\left(\underline{d}_{t:T}, H_{iT}\right) - \eta_t\left(\underline{d}_{t:T}, H_{iT}\right) \right| \right) \\
&+ 2 \sup_{(g_t, \ldots, g_T) \in \mathcal{G}_t \times \cdots \times \mathcal{G}_T} \left| \tilde{Q}_{nt}\left(g_t, \ldots, g_T\right) - \tilde{Q}_t\left(g_t, \ldots, g_T\right) \right|.
\end{aligned}
$$

The first inequality follows from the fact that $\hat{g}^B_{t,e}$ maximizes $\check{Q}_{nt}\left(\cdot, \hat{g}^B_{t+1,e}, \ldots, \hat{g}^B_{T,e}\right)$ over $\mathcal{G}_t$.

Then we have

$$
\begin{aligned}
E_{P^n}\left[\Delta\tilde{Q}^{\dagger}_{t,e}\right] \leq\ & 2E_{P^n}\left[\sup_{(g_t, \ldots, g_T) \in \mathcal{G}_t \times \cdots \times \mathcal{G}_T} \left| \tilde{Q}_{nt}\left(g_t, \ldots, g_T\right) - \tilde{Q}_t\left(g_t, \ldots, g_T\right) \right| \right] \\
&+ \sum_{\underline{d}_{t:T} \in \{0,1\}^{T-t+1}} E_{P^n}\left[\frac{1}{n} \sum_{i=1}^{n} \left| \hat{\eta}_t^{-k}\left(\underline{d}_{t:T}, H_{iT}\right) - \eta_t\left(\underline{d}_{t:T}, H_{iT}\right) \right| \right]
\end{aligned}
$$

Therefore, applying Lemma A.2 to the first term in the right hand side (as in the proof of Lemma A.3 (i)) and Assumption 5.1 (ii) to the second term in the right hand side leads to the result.

(ii) The proof of Lemma F.5 follows from the same argument with the proof of Lemma A.3 (ii).

(iii) We follow the same strategy as in Lemma A.3 (iii). First, note that

$$
\Delta\tilde{Q}_{T,e} = \tilde{Q}_T\left(g^*_T\right) - \tilde{Q}_T\left(\hat{g}^B_{T,e}\right) = \Delta\tilde{Q}^{\dagger}_{T,e}.
$$

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

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

where the inequality follows from Lemma F.5 (ii).

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

$$\begin{aligned}
\Delta\tilde{Q}_{T-k,e} &= \tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_T\right) - \tilde{Q}_{T-k}\left(\hat{g}^B_{T-k,e}, \ldots, \hat{g}^B_{T,e}\right) \\
&= \sum_{s=T-k}^{T}\left[\tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_s, \hat{g}^B_{s+1,e}, \ldots, \hat{g}^B_{T,e}\right) - \tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_{s-1}, \hat{g}^B_{s,e}, \ldots, \hat{g}^B_{T,e}\right)\right] \\
&\le \sum_{s=T-k}^{T}\left[\tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_T\right) - \tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_{s-1}, \hat{g}^B_{s,e}, \ldots, \hat{g}^B_{T,e}\right)\right] \\
&= \sum_{s=T-k+1}^{T}\left[\tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_T\right) - \tilde{Q}_{T-k}\left(g^*_{T-k}, \ldots, g^*_{s-1}, \hat{g}^B_{s,e}, \ldots, \hat{g}^B_{T,e}\right)\right] + \Delta\tilde{Q}^\dagger_{T-k,e} \\
&\le \sum_{s=T-k+1}^{T}\frac{1}{\prod_{\ell=T-k}^{s-1}\kappa_\ell}\Delta\tilde{Q}_{s,e} + \Delta\tilde{Q}^\dagger_{T-k,e},
\end{aligned}$$

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 F.5(ii).

Then, recursively, the following hold:

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

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

$$\begin{aligned}
\Delta\tilde{Q}_{1,e} &\le \Delta\tilde{Q}^\dagger_{1,e} + \sum_{s=1}^{T-1}\frac{2^{T-1-s}}{\prod_{t=1}^{T-s}\kappa_t}\Delta\tilde{Q}^\dagger_{T-s+1,e} \\
&= \Delta\tilde{Q}^\dagger_{1,e} + \sum_{s=1}^{T-1}\frac{2^{s-1}}{\prod_{t=1}^{s}\kappa_t}\Delta\tilde{Q}^\dagger_{s+1,e}.
\end{aligned}$$

□

Proof of Theorem 5.1 (ii). Let $P \in \widetilde{\mathcal{P}}_e \bigcap \mathcal{P}(M, \kappa, \mathcal{G})$ be fixed. By the same argument as in the proof of Theorem 3.6 (ii), it follows for $g^* \in \arg\max_{g \in \mathcal{G}} W(g)$ that

$$\begin{aligned} W_{\mathcal{G}}^* - W(\hat{g}_e^B) &= \tilde{Q}_1(g^*) - \tilde{Q}_1(\hat{g}_e^B) \leq \Delta \tilde{Q}_{1,e} \\ &\leq \Delta \tilde{Q}_{1,e}^\dagger + \sum_{s=1}^{T-1} \frac{2^{s-1}}{\prod_{t=1}^s \kappa_t} \Delta \tilde{Q}_{s+1,e}^\dagger, \end{aligned}$$

where the second inequality follows from Lemma F.5 (iii). Thus, since $W_{\mathcal{G}}^* - W(\hat{g}_e^B) \geq 0$, we have

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

Applying Lemma F.5 (i) to each term in the right hand side gives

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

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

F.4 Proof of Theorem C.1

This section provides the proof of Theorem C.1 for the rank-dependent SWF (A.2). The following is a preliminary lemma.

Lemma F.6. *(Lemma A.5 in Kitagawa and Tetenov (2021)) Let $\mathcal{F}$ be a class of uniformly bounded functions, that is, there exists $\bar{F} < \infty$ such that $||f||_\infty \leq \bar{F}$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph class with VC-dimension $v < \infty$. Let $(Y, Z) \sim P$, where $Y \geq 0$ is a scalar (Y and Z may be dependent). Let $\{(Y_i, Z_i)\}_{i=1}^n \sim P^n$ be an iid sample from P.*

Assume that

$$\int_0^\infty \sqrt{P(Y > y)}dy \leq M.$$

Then, there is a universal constant C such that

$$\int_0^\infty E_{P^n}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)1\{Y_i > y\} - E_P[f(Z)1\{Y > y\}]\right|\right] dy \leq C\bar{F}M\sqrt{\frac{v}{n}}.$$

holds for all $n \geq 1$.

Proof of Theorem C.1. By the same argument as in the proof of Theorem 3.6 (i), it follows that

$$E_{P^n}\left[\sup_{g\in\mathcal{G}} W_\Lambda(g) - W_\Lambda(\hat{g}^S)\right] \leq 2E_{P^n}\left[\sup_{g\in\mathcal{G}}\left|\widehat{W}_\Lambda(g) - W_\Lambda(g)\right|\right]. \tag{A.10}$$

Since $\Lambda(\cdot)$ is convex and nonincreasing,

$$\begin{aligned}
\sup_{g\in\mathcal{G}}\left|\widehat{W}_\Lambda(G) - W_\Lambda(g)\right| &= \sup_{g\in\mathcal{G}}\left|\int_0^\infty \Lambda\left(\widehat{F}_g(y)\vee 0\right)dy - \int_0^\infty \Lambda\left(F_g(y)\right)dy\right| \\
&\leq \sup_{g\in\mathcal{G}}\int_0^\infty \left|\Lambda\left(\widehat{F}_g(y)\vee 0\right) - \Lambda\left(F_g(y)\right)\right|dy \\
&\leq \int_0^\infty \sup_{g\in\mathcal{G}}\left|\Lambda\left(\widehat{F}_g(y)\vee 0\right) - \Lambda\left(F_g(y)\right)\right|dy \\
&\leq |\Lambda'(0)|\int_0^\infty \sup_{g\in\mathcal{G}}\left|\widehat{F}_g(y) - F_g(y)\right|dy.
\end{aligned} \tag{A.11}$$

Combining (A.10) and (A.11) yields

$$E_{P^n}\left[\sup_{g\in\mathcal{G}} W_\Lambda(g) - W_\Lambda(\hat{g}^S)\right] \leq 2\left|\Lambda'(0)\right|\int_0^\infty E_{P^n}\left[\sup_{g\in\mathcal{G}}\left|\widehat{F}_g(y) - F_g(y)\right|\right]dy.$$

For $g \in \mathcal{G}$, let

$$w_g(Z_i) \equiv \frac{\prod_{t=1}^T 1\{g_t(H_{it}) = D_{it}\}}{\prod_{t=1}^T e_t\left(D_{it}, H_{it}\right)}.$$

By Lemma A.1, the class of functions $\mathcal{W} = \{w_g(\cdot) : g \in \mathcal{G}\}$ is a VC-subclass with VC-dimension of at most $\sum_{t=1}^T v_t$. Assumption 2.4 implies that $w_g(Z_i) \in \left[0, 1/\left(\prod_{t=1}^T \kappa_t\right)\right]$;

hence, functions in $\mathcal{W}$ are uniformly bounded by $1/\left(\prod_{t=1}^{T}\kappa_t\right)$.

Since $F_g(y) = 1 - E_P\left[w_g(Z)\cdot 1\left\{\sum_{t=1}^{T} Y_t > y\right\}\right]$, under Assumption 2.1, and $\widehat{F}_g(y) = 1 - \frac{1}{n}\sum_{i=1}^{n} w_g(Z_i)\cdot 1\{\sum_{t=1}^{T} Y_t > y\}$,

$$\left|\widehat{F}_g(y) - F_g(y)\right| = \left|\frac{1}{n}\sum_{i=1}^{n} w_g(Z_i)\cdot 1\left\{\sum_{t=1}^{T} Y_t > y\right\} - E_P\left[w_g(Z)\cdot 1\left\{\sum_{t=1}^{T} Y_t > y\right\}\right]\right|.$$

It follows from Assumption A.3 that

$$\int_0^\infty \sqrt{P\left(\sum_{t=1}^{T} Y_t > y\right)}dy \le \max_{\underline{d}_T\in\{0,1\}^T}\int_0^\infty \sqrt{P\left(\sum_{t=1}^{T} Y_t\left(\underline{d}_t\right) > y\right)}dy \le \Upsilon.$$

Applying Lemma F.6 to (A.11) yields

$$E_{P^n}\left[\sup_{g\in\mathcal{G}} W_\Lambda(g) - W_\Lambda(\hat{g}^S)\right] \le 2C\left|\Lambda'(0)\right|\frac{\Upsilon}{\prod_{t=1}^{T}\kappa_t}\sqrt{\frac{\sum_{t=1}^{T} v_t}{n}}.$$

□

F.5 Proofs of Theorems E.1 and E.3

The following lemma, which directly follows from Lemma 2 in Zhou et al. (2023) and its proof, plays important roles in the proofs of Theorems E.1 and E.3.

Lemma F.7. *Fix $t \in \{1,\ldots,T\}$. For any $\underline{d}_t \in \{0,1\}^t$, let $\{\Gamma_i(\underline{d}_t)\}_{i=1}^n$ be i.i.d. random variables with bounded supports. For any $\underline{g}_{t:T} \in \mathcal{G}_{t:T}$, let $\widetilde{Q}(\underline{g}_{t:T}) \equiv \frac{1}{n}\sum_{i=1}^{n}\Gamma_i^\dagger(\underline{g}_{t:T})$, where $\Gamma_i^\dagger(\underline{g}_{t:T}) \equiv \Gamma_i^\dagger((g_t(H_{it}),\ldots,g_T(H_{iT})))$, and $Q(\underline{g}_{t:T}) \equiv E[\widetilde{Q}(\underline{g}_{t:T})]$. For any $\underline{g}^a_{t:T}, \underline{g}^b_{t:T} \in \mathcal{G}_{t:T}$, denote $\widetilde{\Delta}(\underline{g}^a_{t:T}, \underline{g}^b_{t:T}) \equiv \widetilde{Q}(\underline{g}^a_{t:T}) - \widetilde{Q}(\underline{g}^b_{t:T})$ and $\Delta(\underline{g}^a_{t:T}, \underline{g}^b_{t:T}) \equiv Q(\underline{g}^a_{t:T}) - Q(\underline{g}^b_{t:T})$. Let $\kappa(\cdot)$ denote the entorpy integral defined in Zhou et al. (2023). Then, under Assumption 2.3, the following holds: For any $\delta \in (0,1)$, with probability at least $1-2\delta$,*

$$\sup_{\underline{g}^a_{t:T},\underline{g}^b_{t:T}\in\mathcal{G}_{t:T}}\left|\widetilde{\Delta}(\underline{g}^a_{t:T}, \underline{g}^b_{t:T}) - \Delta(\underline{g}^a_{t:T}, \underline{g}^b_{t:T})\right| \le \left(54.4\sqrt{2}\kappa(\underline{g}_{t:T}) + 435.2 + \sqrt{2\log\frac{1}{\delta}}\right)\sqrt{\frac{V^*_{t:T}}{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where $V^*_{t:T} \equiv \sup_{\underline{g}^a_{t:T}, \underline{g}^b_{t:T} \in \mathcal{G}_{t:T}} E\left[\left(\Gamma^\dagger_i(\underline{g}^a_{t:T}) - \Gamma^\dagger_i(\underline{g}^b_{t:T})\right)^2\right]$.

Proof of Theorem E.1. Let us define

$$\begin{aligned}
\widetilde{V}_i(g) =& \sum_{t=1}^{T} \psi_{it}\left(\underline{g}_t\right) Y_{it} - \sum_{t=1}^{T} \psi_{it}\left(g\right) Q_t^{\underline{g}_{(t+1):T}}\left(H_{it}, D_{i,}\right) \\
&+ \sum_{t=1}^{T} \psi_{i,t-1}\left(\underline{g}_{t-1}\right) Q_t^{\underline{g}_{(t+1):T}}\left(H_{it}, g_t(H_{it})\right), \\
\widehat{V}_i(g) =& \sum_{t=1}^{T} \hat{\psi}_{it}^{-k(i)}\left(\underline{g}_t\right) Y_{it} - \sum_{t=1}^{T} \hat{\psi}_{it}^{-k(i)}\left(g\right) \widehat{Q}_t^{\underline{g}_{(t+1):T}, -k(i)}\left(H_{it}, D_{i,}\right) \\
&+ \sum_{t=1}^{T} \hat{\psi}_{i,t-1}^{-k(i)}\left(\underline{g}_{t-1}\right) \widehat{Q}_t^{\underline{g}_{(t+1):T}, -k(i)}\left(H_{it}, g_t(H_{it})\right),
\end{aligned}$$

with

$$\psi_{it}\left(\underline{g}_t\right) \equiv \frac{\prod_{s=1}^{t} 1\{D_{is} = g_s(H_{is})\}}{\prod_{s=1}^{t} e_t(H_{is}, g_s(H_{is}))}.$$

We also define $V(g) = E\left[\widetilde{V}_i(g)\right]$. Lemma B.2 in Sakaguchi (2024) shows that $V(g) = W(g)$ under Assumptions 2.1.

For any $g^a, g^b \in \mathcal{G}$, we define

$$\begin{aligned}
\Delta(g^a, g^b) &\equiv V(g^a) - V(g^b), \\
\widetilde{\Delta}(g^a, g^b) &\equiv \frac{1}{n}\sum_{i=1}^{n} \widetilde{V}_i(g^a) - \frac{1}{n}\sum_{i=1}^{n} \widetilde{V}_i(g^b), \\
\widehat{\Delta}(g^a, g^b) &\equiv \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_i(g^a) - \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_i(g^b).
\end{aligned}$$

Let $g^*_{opt} \in \arg\max_{g \in \mathcal{G}} W(g)$. A standard argument of the statistical learning theory gives

$$\begin{aligned}
W^*_{\mathcal{G}} - W\left(\hat{g}^{AIPW}\right) &= \Delta(g^*_{opt}, \hat{g}^{AIPW}) \\
&\leq \Delta(g^*_{opt}, \hat{g}^{AIPW}) - \widehat{\Delta}(g^*_{opt}, \hat{g}^{AIPW}) \\
&\leq \sup_{g^a, g^b \in \mathcal{G}} \left|\Delta(g^a, g^b) - \widehat{\Delta}(g^a, g^b)\right| \\
&\leq \sup_{g^a, g^b \in \mathcal{G}} \left|\Delta(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right| + \sup_{g^a, g^b \in \mathcal{G}} \left|\widehat{\Delta}(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right|,
\end{aligned} \tag{A.12}$$

where the first inequality follows because $\hat{g}^{AIPW}$ maximizes $\widehat{V}(g)$ over $\mathcal{G}$; hence, $\widehat{\Delta}(g^*, \hat{g}^{AIPW}) \leq 0$.

We can now evaluate $W^*_{\mathcal{G}} - W(\hat{g}^{AIPW})$ through evaluating $\sup_{g^a, g^b \in \mathcal{G}} \left|\Delta(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right|$ and $\sup_{g^a, g^b \in \mathcal{G}} \left|\widehat{\Delta}(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right|$. As for the former, under Assumptions 2.1–2.4, we can apply Lemma F.7 to obtain the following result: For any stage t and $\delta \in (0, 1)$, with probability at least $1 - 2\delta$,

$$\sup_{g^a, g^b \in \mathcal{G}} \left|\Delta(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right| \leq \left(54.4\sqrt{2}\kappa(\mathcal{G}) + 435.2 + \sqrt{2\log\frac{1}{\delta}}\right)\sqrt{\frac{V^*}{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{A.13}$$

where $V^* \equiv \sup_{g^a, g^b \in \mathcal{G}} E\left[\left(\widetilde{V}_i(g^a) - \widetilde{V}_i(g^b)\right)^2\right] < \infty$. Note that $\kappa(\mathcal{G}) \leq 2.5\sqrt{VC(\mathcal{G})} < \infty$ from Remark 8 in Zhou et al. (2023) and Assumption 2.3.

As for the latter $\sup_{g^a, g^b \in \mathcal{G}} \left|\widehat{\Delta}(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right|$, under Assumptions 2.1–2.4 and E.1, we can obtain from Lemma A.3 in Sakaguchi (2024) that

$$\sup_{g^a, g^b \in \mathcal{G}} \left|\widehat{\Delta}(g^a, g^b) - \widetilde{\Delta}(g^a, g^b)\right| = O_p\left(n^{-\min\{1/2, \tau\}}\right). \tag{A.14}$$

Combining (A.12), (A.13), and (A.14) leads to the result. □

Proof of Lemma E.2. Under Assumption E.2 and the redefinition $H_t(\underline{d}_t) = (\underline{D}_{t-1}, \underline{X}_t)$,

$$\widetilde{Y}_t\left(\underline{g}_t\right) = \sum_{\underline{d}_t \in \{0,1\}^t} Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\left\{g_s\left(\underline{d}_{s-1}, \underline{X}_s\right) = d_s\right\},$$

where we use the fact that $\underline{X}_t = \underline{X}_t(\underline{d}_t)$ for any $\underline{d}_t$. Then we have

$$W_t\left(\underline{g}_t\right) = \sum_{\underline{d}_t \in \{0,1\}^t} E_P\left[\gamma_t Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\left\{g_s\left(\underline{d}_{s-1}, \underline{X}_s\right) = d_s\right\}\right].$$

It follows that

$$\sum_{\underline{d}_t \in \{0,1\}^t} E_P\left[\mu_t(\underline{d}_t, \underline{X}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\}\right]$$

$$= \sum_{\underline{d}_t \in \{0,1\}^t} E_P \left[E_P \left[\gamma_t Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\} \middle| \underline{D}_t = \underline{d}_t, \underline{X}_t \right] \right]$$
$$= \sum_{\underline{d}_t \in \{0,1\}^t} E_P \left[E_P \left[\gamma_t Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\} \middle| \underline{D}_{t-1} = \underline{d}_{t-1}, \underline{X}_t \right] \right]$$
$$= \sum_{\underline{d}_t \in \{0,1\}^t} E_P \left[E_P \left[\gamma_t Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\} \middle| \underline{D}_{t-1} = \underline{d}_{t-1}, \underline{X}_{t-1} \right] \right].$$

where the second equality follows from Assumption 2.1 and the third follows by the law of iterated expectations. Applying the same argument recursively,

$$\sum_{\underline{d}_t \in \{0,1\}^t} E_P \left[\mu_t(\underline{d}_t, \underline{X}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\} \right]$$
$$= \sum_{\underline{d}_t \in \{0,1\}^t} E_P \left[\gamma_t Y_t(\underline{d}_t) \cdot \prod_{s=1}^{t} 1\{d_s = g_s(\underline{d}_{s-1}, \underline{X}_s)\} \right] = W_t\left(\underline{g}_t\right).$$

□

Proof of Theorem E.3. For any $g \in \mathcal{G}$, let $\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_T)$, where $\tilde{g}_t : \underline{\mathcal{X}}_t \to \{0,1\}$, be recursively defined as

$$\begin{aligned}
\tilde{g}_1(x_1) &= g_1(x_1), \\
\tilde{g}_2(\underline{x}_2) &= g_2(\tilde{g}_1(x_1), \underline{x}_2), \\
\tilde{g}_3(\underline{x}_3) &= g_2(\tilde{g}_1(x_1), \tilde{g}_2(\underline{x}_2), \underline{x}_3), \\
&\vdots \\
\tilde{g}_t(\underline{x}_t) &= g_t(\tilde{g}_1(x_1), \tilde{g}_2(\underline{x}_2), \ldots, \tilde{g}_{t-1}(\underline{x}_{t-1}), \underline{x}_t), \\
&\vdots \\
\tilde{g}_T(\underline{x}_T) &= g_T(\tilde{g}_1(x_1), \tilde{g}_2(\underline{x}_2), \ldots, \tilde{g}_{T-1}(\underline{x}_{T-1}), \underline{x}_T).
\end{aligned}$$

For the class $\mathcal{G}$ of DTRs, we define the class of $\tilde{g}$ as

$$\widetilde{\mathcal{G}} \equiv \{\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_T) : g \in \mathcal{G}\}.$$

Let $\widetilde{\mathcal{G}}_{1:t} := \widetilde{\mathcal{G}}_1 \times \cdots \times \widetilde{\mathcal{G}}_t$. We denote by $\kappa(\cdot)$ the entropy integral defined in Zhou et al.

(2023). Note that $\kappa(\widetilde{\mathcal{G}}) \leq \kappa(\mathcal{G})$.

Without loss of generality, we suppose that $\gamma_1 = \gamma_2 = \cdots = \gamma_T$. Let us define, for $t = 1, \ldots, T$,

$$\Gamma_{it}(\underline{d}_t) \equiv \frac{Y_{it} - \mu_t(\underline{d}_t, \underline{X}_{it})}{\eta_t(\underline{d}_t, \underline{X}_{it})} \cdot 1\{\underline{D}_{it} = \underline{d}_t\} + \mu_t(\underline{d}_t, \underline{X}_{it}),$$
$$\widehat{\Gamma}_{it}(\underline{d}_t) \equiv \frac{Y_{it} - \hat{\mu}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it})}{\hat{\eta}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it})} \cdot 1\{\underline{D}_{it} = \underline{d}_t\} + \hat{\mu}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it}),$$

where $\eta_t(\underline{d}_t, \underline{x}_t) \equiv \prod_{s=1}^t e_s(d_s, H_{is})$ and $\hat{\eta}_t^{-k(i)}(\underline{d}_t, \underline{x}_t) \equiv \prod_{s=1}^t \hat{e}_s^{-k(i)}(d_s, H_{is})$.

Let $\underline{\tilde{g}}_t(\underline{x}_t) = (\tilde{g}_1(x_1), \tilde{g}_2(\underline{x}_2), \ldots, \tilde{g}_t(\underline{x}_t))$. Given a fixed DTR $\tilde{g} \in \widetilde{\mathcal{G}}$ constructed from $g \in \mathcal{G}$, with some abuse of the notation, we define for $t = 1, \ldots, T$,

$$\Gamma_{it}(\underline{\tilde{g}}_t) \equiv \sum_{\underline{d}_t \in \{0,1\}^t} \Gamma_{it}(\underline{d}_t) \cdot 1\{\underline{\tilde{g}}_t(\underline{x}_t) = \underline{d}_t\},$$
$$\widehat{\Gamma}_{it}(\underline{\tilde{g}}_t) \equiv \sum_{\underline{d}_t \in \{0,1\}^t} \widehat{\Gamma}_{it}(\underline{d}_t) \cdot 1\{\underline{\tilde{g}}_t(\underline{x}_t) = \underline{d}_t\}.$$

Note that $(1/n)\sum_{t=1}^t \sum_{i=1}^n \widehat{\Gamma}_{it}(\underline{\tilde{g}}_t) = \widehat{W}^{DR}(g)$. Hence,

$$\max_{\tilde{g} \in \widetilde{\mathcal{G}}} \frac{1}{n} \sum_{t=1}^t \sum_{i=1}^n \widehat{\Gamma}_{it}(\underline{\tilde{g}}_t) = \max_{g \in \mathcal{G}} \widehat{W}^{DR}(g).$$

Note also that $(1/n)\sum_{i=1}^n \Gamma_{it}\left(\underline{\tilde{g}}_t\right)$ is an oracle estimate of $W_t(\underline{g}_t)$ with oracle access to $\{\mu_s(\cdot,\cdot) : s = 1, \ldots, t\}$ and $\{e_s(\cdot,\cdot) : s = 1, \ldots, t\}$. For $\underline{g}_t \in \mathcal{G}_{1:t}$, we define $\widetilde{W}_t(\underline{\tilde{g}}_t) \equiv E\left[\Gamma_{it}(\underline{\tilde{g}}_t)\right]$. Note that $W_t(\underline{g}_t) = \widetilde{W}_t(\underline{\tilde{g}}_t)$ by Lemma E.2 under Assumptions 2.1 and E.2.

Following the analysis of Zhou et al. (2023), we define the policy value difference function $\Delta_t(\cdot;\cdot) : \widetilde{\mathcal{G}}_{1:t} \times \widetilde{\mathcal{G}}_{1:t} \to \mathbb{R}$, the oracle influence difference function $\widetilde{\Delta}_t(\cdot;\cdot) : \widetilde{\mathcal{G}}_{1:t} \times \widetilde{\mathcal{G}}_{1:t} \to \mathbb{R}$, and the estimated policy value difference function $\widehat{\Delta}_t(\cdot;\cdot) : \widetilde{\mathcal{G}}_{1:t} \times \widetilde{\mathcal{G}}_{1:t} \to \mathbb{R}$, respectively, as follows: For $\underline{\tilde{g}}_t^a = (\tilde{g}_1^a, \ldots, \tilde{g}_t^a) \in \widetilde{\mathcal{G}}_{1:t}$ and $\underline{\tilde{g}}_t^b = (\tilde{g}_1^b, \ldots, \tilde{g}_t^b) \in \widetilde{\mathcal{G}}_{1:t}$,

$$\Delta_t(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b) \equiv \widetilde{W}_t(\underline{\tilde{g}}_t^a) - \widetilde{W}_t(\underline{\tilde{g}}_t^b),$$
$$\widetilde{\Delta}_t(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^n \Gamma_{it}\left(\underline{\tilde{g}}_t^a\right) - \frac{1}{n}\sum_{i=1}^n \Gamma_{it}\left(\underline{\tilde{g}}_t^b\right),$$

$$\widehat{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^n \widehat{\Gamma}_{it}\left(\underline{\tilde{g}}_t^a\right) - \frac{1}{n}\sum_{i=1}^n \widehat{\Gamma}_{it}\left(\underline{\tilde{g}}_t^b\right).$$

Note that $\widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)$ is an unbiased estimator of the policy value difference function $\Delta_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)$. From the definitions,

$$W_{\mathcal{G}}^* - W(\hat{g}^{DR}) = \sum_{t=1}^T \Delta_t\left(\underline{\tilde{g}}_{1:t}^*;\underline{\tilde{\hat{g}}}_{1:t}^{DR}\right).$$

A standard argument of the statistical learning theory gives

$$\begin{aligned}
W_{\mathcal{G}}^* - W(\hat{g}^{DR}) &= \sum_{t=1}^T \Delta_t\left(\underline{\tilde{g}}_{1:t}^*;\underline{\tilde{\hat{g}}}_{1:t}^{DR}\right) \\
&\leq \sum_{t=1}^T \Delta_t\left(\underline{\tilde{g}}_{1:t}^*;\underline{\tilde{\hat{g}}}_{1:t}^{DR}\right) - \sum_{t=1}^T \widehat{\Delta}_t\left(\underline{\tilde{g}}_{1:t}^*;\underline{\tilde{\hat{g}}}_{1:t}^{DR}\right) \\
&\leq \sum_{t=1}^T \sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} |\Delta_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \widehat{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)| \\
&\leq \sum_{t=1}^T \sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} |\Delta_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)| \\
&\quad + \sum_{t=1}^T \sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} |\widehat{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)|, \qquad \text{(A.15)}
\end{aligned}$$

where the first inequality follows because $\tilde{\hat{g}}^{DR}$ maximizes $(1/n)\sum_{t=1}^T\sum_{i=1}^n \widehat{\Gamma}_{it}\left(\underline{\tilde{g}}_{1:t}\right)$ over $\widetilde{\mathcal{G}}$; hence, $\sum_{t=1}^T \widehat{\Delta}_t\left(\underline{\tilde{g}}_{1:t}^*;\underline{\tilde{\hat{g}}}_{1:t}^{DR}\right) \leq 0$.

We can now evaluate $W_{\mathcal{G}}^* - W(\hat{g}^{DR})$ through evaluating $\sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} |\Delta_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)|$ and $\sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} |\widehat{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)|$ for each t. As for the former, under Assumptions 2.1–2.4, we can apply Lemma F.7 for the oracle influence difference function with some modifications to the notations to obtain the following result: For any stage t and $\delta \in (0,1)$, with probability at least $1-2\delta$,

$$\begin{aligned}
\sup_{\underline{\tilde{g}}_t^a,\underline{\tilde{g}}_t^b\in\widetilde{\mathcal{G}}_{1:t}} \left|\widetilde{\Delta}_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b) - \Delta_t(\underline{\tilde{g}}_t^a;\underline{\tilde{g}}_t^b)\right| &\leq \left(54.4\sqrt{2}\kappa(\widetilde{\mathcal{G}}_{1:t}) + 435.2 + \sqrt{2\log\frac{1}{\delta}}\right)\sqrt{\frac{V_t^*}{n}} \\
&\quad + o\left(\frac{1}{\sqrt{n}}\right), \qquad \text{(A.16)}
\end{aligned}$$

where

$$V_t^* \equiv \sup_{\tilde{\underline{g}}_t^a, \tilde{\underline{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} E\left[\left(\Gamma_{it}(\tilde{\underline{g}}_t^a) - \Gamma_{it}(\tilde{\underline{g}}_t^b)\right)^2\right] < \infty.$$

Note that from Remark 8 in Zhou et al. (2023), $\kappa(\widetilde{\mathcal{G}}_{1:t}) \leq 2.5\sqrt{VC(\widetilde{\mathcal{G}}_{1:t})}$. Using Lemma A.1 and the fact that $VC(\widetilde{\mathcal{G}}_{1:t}) \leq VC(\mathcal{G}_{1:t})$, it follwos that

$$\kappa(\widetilde{\mathcal{G}}_{1:t}) \leq 2.5\sqrt{VC(\mathcal{G}_{1:t})} \leq 2.5\sum_{s=1}^{t} v_s < \infty, \tag{A.17}$$

where the last inequality follows from Assumption 2.3.

We next consider evaluating $\sup_{\tilde{\underline{g}}_t^a, \tilde{\underline{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} |\widehat{\Delta}_t(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) - \widetilde{\Delta}_t(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b)|$. We employ the general strategy of the poof of Zhou et al. (2023, Lemma 3). Fix t. For any $\underline{d}_t \in \{0,1\}^t$, let

$$\widetilde{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^{n} \Gamma_{it}^{\underline{d}_t}\left(\tilde{\underline{g}}_t^a\right) - \frac{1}{n}\sum_{i=1}^{n} \Gamma_{it}^{\underline{d}_t}\left(\tilde{\underline{g}}_t^b\right),$$
$$\widehat{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_{it}^{\underline{d}_t}\left(\tilde{\underline{g}}_t^a\right) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_{it}^{\underline{d}_t}\left(\tilde{\underline{g}}_t^b\right),$$

with $\Gamma_{it}^{\underline{d}_t}(\tilde{\underline{g}}_t) \equiv \Gamma_{it}(\underline{d}_t) \cdot 1\{\tilde{\underline{g}}_t(\underline{x}_t) = \underline{d}_t\}$ and $\widehat{\Gamma}_{it}^{\underline{d}_t}(\tilde{\underline{g}}_t) \equiv \sum_{\underline{d}_t \in \{0,1\}^t} \widehat{\Gamma}_{it}(\underline{d}_t) \cdot 1\{\tilde{\underline{g}}_t(\underline{x}_t) = \underline{d}_t\}$. Noting that $\widehat{\Delta}_t(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) - \widetilde{\Delta}_t(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) = \sum_{\underline{d}_t \in \{0,1\}^t}\left(\widehat{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) - \widetilde{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b)\right)$, we will provide an upper bound for each $\widehat{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) - \widetilde{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b)$. To do so, we make the following decomposition:

$$\widehat{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) - \widetilde{\Delta}_t^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) = S_{1,t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) + S_{2,t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) + S_{3,t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b),$$

where

$$S_{(A1),t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^{n} G_{i,\tilde{\underline{g}}_t^a,\tilde{\underline{g}}_t^b}^{\underline{d}_t}\left(\hat{\mu}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it}) - \mu_t^{-k(i)}(\underline{d}_t, \underline{X}_{it})\right)\left(1 - \frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\eta_t(\underline{d}_t, \underline{X}_{it})}\right),$$
$$S_{(A2),t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^{n} G_{i,\tilde{\underline{g}}_t^a,\tilde{\underline{g}}_t^b}^{\underline{d}_t}\left(Y_{it} - \mu_t^{-k(i)}(\underline{d}_t, \underline{X}_{it})\right)\left(\frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\hat{\eta}_t^{-k}(\underline{d}_t, \underline{X}_{it})} - \frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\eta_t(\underline{d}_t, \underline{X}_{it})}\right),$$
$$S_{(A3),t}^{\underline{d}_t}(\tilde{\underline{g}}_t^a; \tilde{\underline{g}}_t^b) \equiv \frac{1}{n}\sum_{i=1}^{n} G_{i,\tilde{\underline{g}}_t^a,\tilde{\underline{g}}_t^b}^{\underline{d}_t}\left(\mu_t^{-k(i)}(\underline{d}_t, \underline{X}_{it}) - \hat{\mu}_t^{-k(i)}(\underline{d}_t, \underline{X}_{it})\right)\left(\frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\hat{\eta}_t^{-k}(\underline{d}_t, \underline{X}_{it})} - \frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\eta_t(\underline{d}_t, \underline{X}_{it})}\right),$$

with $G_{i,\tilde{\underline{g}}_t^a,\tilde{\underline{g}}_t^b}^{\underline{d}_t} := 1\{\tilde{\underline{g}}_t^a(\underline{X}_{it}) = \underline{d}_t\} - 1\{\tilde{\underline{g}}_t^b(\underline{X}_{it}) = \underline{d}_t\}$.

For each fold k, define

$$S^{\underline{d}_t,k}_{(A1),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t) \equiv \frac{1}{n}\sum_{\{i|k(i)=k\}} G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right);$$
$$S^{\underline{d}_t,k}_{(A2),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t) \equiv \frac{1}{n}\sum_{\{i|k(i)=k\}} G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t}\left(Y_{it}-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\hat{\eta}^{-k}_t(\underline{d}_t,\underline{X}_{it})}-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right).$$

Note that $S^{\underline{d}_t}_{(A1),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t)=\sum_{k=1}^K S^{\underline{d}_t,k}_{(A1),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t)$ and $S^{\underline{d}_t}_{(A2),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t)=\sum_{k=1}^K S^{\underline{d}_t,k}_{(A2),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t)$.

Fix $k\in\{1,\ldots,K\}$. We first consider $S^{\underline{d}_t}_{(A1),t}(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t)$. Since $\hat{\mu}^{-k}_t(\underline{d}_t,\cdot)$ is computed using the data in the rest $K-1$ folds, when the data $\{Z_i : k(i)\neq k\}$ in the rest $K-1$ folds is conditioned, $\hat{\mu}^{-k}_t(\underline{d}_t,\cdot)$ is fixed; hence, $\widetilde{S}^{\underline{d}_t,k}_{(A1),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)$ is a sum of i.i.d. bounded random variables under Assumptions 2.2, 2.4, and E.3 (ii).

It follows that

$$\begin{aligned}
&E\left[G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\middle|\hat{\mu}^{-k}_t(\underline{d}_t,\cdot)\right]\\
=&E\left[G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)E\left[\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\middle|\underline{X}_{it}\right]\middle|\hat{\mu}^{-k}_t(\underline{d}_t,\cdot)\right]\\
=&0,
\end{aligned}$$

where the last line follows from the sequential independence assumption (Assumption 2.1). Hence, $\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|,k\widetilde{S}^{\underline{d}_t}_{(A1),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)\right|$ can be written as

$$\begin{aligned}
&\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\widetilde{S}^{\underline{d}_t,k}_{(A1),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)\right|\\
=&\frac{1}{K}\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\frac{1}{n/K}\sum_{i\in I_k}\left\{G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\right.\right.\\
&\left.\left.-E\left[G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\middle|\hat{\mu}^{-k}_t(\underline{d}_t,\cdot)\right]\right\}\right|.
\end{aligned}$$

By applying Lemma F.7 with setting $i\in I_k$ and

$$\Gamma_i(d_t)=G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t}\left(\hat{\mu}^{-k}_t(\underline{d}_t,\underline{X}_{it})-\mu^{-k}_t(\underline{d}_t,\underline{X}_{it})\right)\left(1-\frac{1\{\underline{D}_{it}=\underline{d}_t\}}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right),$$

the following holds: $\forall \delta > 0$, with probability at least $1 - 2\delta$,

$$
\begin{aligned}
&\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t, k}_{(A1),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| \\
&\leq o\left(n^{-1/2}\right) + \left(54.4\kappa \left(\widetilde{\mathcal{G}}_{1:t} \right) + 435.2 + \sqrt{2\log(1/\delta)} \right) \\
&\times \left[\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} E\left[\left(G^{\underline{d}_t}_{i, \underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b} \right)^2 \left(\hat{\mu}_t^{-k}(\underline{d}_t, \underline{X}_{it}) - \mu_t^{-k}(\underline{d}_t, \underline{X}_{it}) \right)^2 \right. \right. \\
&\left. \left. \times \left(1 - \frac{1\{\underline{D}_{it} = \underline{d}_t\}}{\eta_t(\underline{d}_t, \underline{X}_{it})} \right) \middle| \hat{\mu}_t^{-k}\left(\underline{d}_t, \cdot\right) \right] \middle/ \left(\frac{n}{K} \right) \right]^{1/2} \\
&\leq o\left(n^{-1/2}\right) + \sqrt{K} \cdot \left(54.4\kappa \left(\widetilde{\mathcal{G}}_{1:t} \right) + 435.2 + \sqrt{2\log(1/\delta)} \right) \cdot \left(1 - \frac{1}{\eta} \right)^t \\
&\times \sqrt{\frac{E\left[\left(\hat{\mu}_t^{-k}(\underline{d}_t, \underline{X}_{it}) - \mu_t^{-k}(\underline{d}_t, \underline{X}_{it}) \right)^2 \middle| \hat{\mu}_t^{-k}\left(\underline{d}_t, \cdot\right) \right]}{n}},
\end{aligned}
$$

where the last inequality follows from $\left(G^{\underline{d}_t}_{i, \underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b} \right)^2 \leq 1$ a.s. and Assumption 2.4 (overlap condition). From Assumptions 2.2 and E.3 (ii), we have $E\left[\left(\hat{\mu}_t^{-k}(\underline{d}_t, \underline{X}_{it}) - \mu_t^{-k}(\underline{d}_t, \underline{X}_{it}) \right)^2 \right] < \infty$. Hence, Markov's inequality leads to

$$
E\left[\left(\hat{\mu}_t^{-k}(\underline{d}_t, \underline{X}_{it}) - \mu_t^{-k}(\underline{d}_t, \underline{X}_{it}) \right)^2 \middle| \hat{\mu}_t^{-k}\left(\underline{d}_t, \cdot\right) \right] = O_p(1).
$$

Note also that $\kappa(\widetilde{\mathcal{G}}_{1:t}) < \infty$ from (A.17). Combining these results, we have

$$
\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t, k}_{(A1),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| = O_p\left(\frac{1}{\sqrt{n}} \right).
$$

Consequently,

$$
\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t}_{(A1),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| \leq \sum_{k=1}^{K} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t, k}_{(A1),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| = O_p\left(\frac{1}{\sqrt{n}} \right). \tag{A.18}
$$

By the same argument, we have $\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t, k}_{(A2),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| = O_p\left(1/\sqrt{n}\right)$. Hence

$$
\sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t}_{(A2),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| \leq \sum_{k=1}^{K} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}^{\underline{d}_t, k}_{(A2),t} \left(\underline{\tilde{g}}_t^a ; \underline{\tilde{g}}_t^b \right) \right| = O_p\left(\frac{1}{\sqrt{n}} \right). \tag{A.19}
$$

We next consider to bound $S^{\underline{d}_t}_{(A3),t}(\cdot,\cdot)$ from above. It follows that

$$
\begin{aligned}
&\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\widetilde{S}^{\underline{d}_t}_{(A3),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)\right| \\
&= \frac{1}{n}\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\sum_{i=1}^{n} G^{\underline{d}_t}_{i,\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t}\left(\mu_t(\underline{d}_t,\underline{X}_{it})-\hat{\mu}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})\right)\left(\frac{1}{\hat{\eta}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})}-\frac{1}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\right| \\
&\leq \frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}\left|\left(\mu_t(\underline{d}_t,\underline{X}_{it})-\hat{\mu}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})\right)\right|\cdot\left|\left(\frac{1}{\hat{\eta}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})}-\frac{1}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}\left(\mu_t(\underline{d}_t,\underline{X}_{it})-\hat{\mu}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})\right)^2}\sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}\left(\frac{1}{\hat{\eta}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})}-\frac{1}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)^2},
\end{aligned}
$$

where the last inequality follows from Cauchy-Schwartz inequality. Taking expectation of both sides yields:

$$
\begin{aligned}
E\left[\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\widetilde{S}^{\underline{d}_t}_{(A3),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)\right|\right] &\leq E\left[\sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}\left(\mu_t(\underline{d}_t,\underline{X}_{it})-\hat{\mu}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})\right)^2}\right] \\
&\quad\times E\left[\sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}\left(\frac{1}{\hat{\eta}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})}-\frac{1}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)^2}\right] \\
&\leq \sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}E\left[\left(\mu_t(\underline{d}_t,\underline{X}_{it})-\hat{\mu}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})\right)^2\right]} \\
&\quad\times\sqrt{\frac{1}{n}\sum_{\{i|\underline{D}_{it}=\underline{d}_t\}}E\left[\left(\frac{1}{\hat{\eta}_t^{-k(i)}(\underline{d}_t,\underline{X}_{it})}-\frac{1}{\eta_t(\underline{d}_t,\underline{X}_{it})}\right)^2\right]} \\
&= o(n^{-\tau'/2}),
\end{aligned}
$$

where the second inequality follows from Cauchy-Schwartz inequality and the last line follows from Assumption E.3 (i). Then applying Markov's inequality leads to

$$
\sup_{\tilde{\underline{g}}^a_t,\tilde{\underline{g}}^b_t\in\widetilde{\mathcal{G}}_{1:t}}\left|\widetilde{S}^{\underline{d}_t}_{(A3),t}\left(\tilde{\underline{g}}^a_t;\tilde{\underline{g}}^b_t\right)\right| = O_P\left(n^{-\tau'/2}\right). \tag{A.20}
$$

We therefore obtain

$$
\begin{aligned}
&\sum_{t=1}^{T} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} |\widehat{\Delta}_t(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b)| \\
&\leq \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} |\widehat{\Delta}_t^{\underline{d}_t}(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b) - \widetilde{\Delta}_t^{\underline{d}_t}(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b)| \\
&\leq \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}_{(A1),t}^{\underline{d}_t} \left(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b \right) \right| + \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}_{(A2),t}^{\underline{d}_t} \left(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b \right) \right| \\
&+ \sum_{t=1}^{T} \sum_{\underline{d}_t \in \{0,1\}^t} \sup_{\underline{\tilde{g}}_t^a, \underline{\tilde{g}}_t^b \in \widetilde{\mathcal{G}}_{1:t}} \left| \widetilde{S}_{(A3),t}^{\underline{d}_t} \left(\underline{\tilde{g}}_t^a; \underline{\tilde{g}}_t^b \right) \right| \\
&= O_p \left(n^{-\min\{1/2, \tau'/2\}} \right), \qquad \text{(A.21)}
\end{aligned}
$$

where the last line follows from (A.18), (A.19), and (A.20).

Combining (A.15), (A.16), (A.17), and (A.21) leads to the result (A.7). □

G Additional Simulation Results

We conduct an additional simulation study to examine the finite sample performance for the estimation methods proposed in Section 3 under the circumstance that the sequential independence assumption does not hold due to the presence of unobserved heterogeneity. We consider the same DGPs as those used in Section 6, except that the treatment assignments D_1 and D_2 are distributed as

$$
D_1 \sim 1\{N(0,1) + \rho \cdot U_1 \geq 0\} \text{ and } D_2 \sim 1\{N(0,1) + \rho \cdot U_2 \geq 0\}. \qquad \text{(A.22)}
$$

Recall that the potential outcomes $Y_1(d_1)$ and $Y_2(d_1, d_2)$ depend on U_1 and U_2, respectively. Hence, unless $\rho \neq 0$, the sequential independence assumption (Assumption 2.1) is not satisfied. We consider two values of ρ: $\rho = -1$ and 1. For each $j = 1, 2, 3$, we label the DGP that is the same as DGP j used in Section 6 except for that D_1 and D_2 follow equation (A.22) with $\rho = -1$ and 1 as DGPs j' and j'', respectively.

Table G.1 presents the results of 500 simulations with sample sizes $n = 200$, 500, and 800, where we compare Q-learning, backward DEWM, and simultaneous DEWM and calculate the mean and median welfare achieved by each estimated DTR. Panel (A) of

Table G.1 shows that in the case of $\rho = -1$, simultaneous DEWM leads to the lower mean welfare in DGP 3′ than Q-learning and backward DEWM. This result differs from the simulation results in Section 6, where the DGPs satisfy the sequential independence assumption, and simultaneous DEWM leads to the highest mean welfare in DGP3. Panel (B) of Table G.1 shows that the DGPs 1″, 2″, and 3″ lead to similar results to those with DGPs 1-3 in terms of the order of mean/median welfare among the three methods.

Table G.1: Additional Monte Carlo Simulation Results

Panel(A) DGPs 1′-3′

		n=200			n=500			n=800		
	DGP	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Q-learning	1′	1.847	1.850	0.045	1.855	1.852	0.036	1.859	1.861	0.039
B-DEWM	1′	1.614	1.654	0.201	1.650	1.677	0.166	1.658	1.677	0.169
S-DEWM	1′	1.424	1.499	0.276	1.459	1.544	0.267	1.493	1.577	0.273
Q-learning	2′	3.120	3.123	0.059	3.123	3.126	0.057	3.126	3.128	0.059
B-DEWM	2′	2.524	2.589	0.520	2.669	2.816	0.437	2.701	2.816	0.400
S-DEWM	2′	2.546	2.784	0.609	2.834	3.010	0.471	2.943	3.066	0.412
Q-learning	3′	1.578	1.573	0.227	1.555	1.549	0.187	1.531	1.530	0.169
B-DEWM	3′	1.703	1.755	0.176	1.719	1.743	0.138	1.716	1.729	0.136
S-DEWM	3′	1.358	1.341	0.156	1.356	1.341	0.118	1.353	1.342	0.099

Panel(B) DGPs 1″-3″

		n=200			n=500			n=800		
	DGP	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Q-learning	1″	3.097	3.098	0.036	3.102	3.100	0.035	3.101	3.101	0.035
B-DEWM	1″	2.849	2.960	0.296	3.000	3.059	0.173	3.004	3.069	0.200
S-DEWM	1″	2.877	3.002	0.335	3.005	3.062	0.227	3.041	3.075	0.191
Q-learning	2″	5.200	5.201	0.059	5.210	5.208	0.060	5.208	5.210	0.060
B-DEWM	2″	4.595	4.955	0.775	4.865	5.102	0.596	4.891	5.119	0.645
S-DEWM	2″	4.847	5.014	0.540	5.037	5.099	0.305	5.070	5.135	0.385
Q-learning	3″	2.193	2.192	0.114	2.184	2.180	0.109	2.191	2.185	0.105
B-DEWM	3″	2.021	1.919	0.276	2.194	2.262	0.250	2.235	2.259	0.199
S-DEWM	3″	2.299	2.321	0.166	2.296	2.304	0.142	2.309	2.320	0.148

Note: Mean and Median represent the mean and median of the population mean welfares achieved by the estimated DTRs across the simulations; 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 mean the Backward and Simultaneous DEWM methods, respectively.

H Computation

In this appendix, we explain computation of the backward and simultaneous DEWM with $\mathcal{G}_t$ $(t = 1, \ldots, T)$ being classes of the linear treatment rules. The non-convexity of the objective functions make these computations challenging. However, the optimization problems can be formulated as Mixed Integer Linear Programming (MILP) problems, for which some efficient softwares (e.g., CPLEX; Gurobi) are available. In the following subsections, we illustrate the MILP formalization for each of the backward and simultaneous DEWM in the case of $T = 2$. We suppose that the class of feasible treatment rules for each stage $t = 1, 2$ takes the form of $\mathcal{G}_t = \left\{1\left\{(1, H_t')\beta_t \geq 0\right\} : \beta_t \in \mathcal{B}_t \subset \mathbb{R}^{(k+2)t-1}\right\}$ where $\mathcal{B}_t$ is a compact set.

H.1 Backward DEWM

Using slightly different notation from Section 3.1, the first step of the backward DEWM method is

$$\max_{g_2 \in \mathcal{G}_2} \sum_{i=1}^{n} m_{i2}^B g_2,$$

where

$$m_{i2}^B = \left(\frac{D_{i2}}{e_2\left(1, H_{i2}\right)} - \frac{1 - D_{i2}}{e_2\left(0, H_{i2}\right)} \right) \gamma_2 Y_{i2}.$$

Let $\hat{g}_2^B$ be a maximizer of the above problem. Then, the second step of the backward DEWM method is

$$\max_{g_1 \in \mathcal{G}_1} \sum_{i=1}^{n} m_{i1}^B g_1,$$

where

$$\begin{aligned} m_{i1}^B = & \left(\frac{D_{i1}}{e_1\left(1, H_{i1}\right)} - \frac{1 - D_{i1}}{e_2\left(0, H_{i1}\right)} \right) \\ & \times \left(\frac{D_{i2} \hat{g}_2^B\left(H_{i2}\right)}{e_2\left(1, H_{i2}\right)} - \frac{\left(1 - D_{i2}\right)\left(1 - \hat{g}_2^B\left(H_{i2}\right)\right)}{e_2\left(0, H_{i2}\right)} \right) \left(\gamma_1 Y_{1i} + \gamma_2 Y_{i2}\right). \end{aligned}$$

When the class of DTRs is constrained to the class of linear eligibility rules, each step of the backward DEWM method described in Section 3.1 can be formulated as MILP problem. The optimization problem in the first step is equivalent to the following MILP problem:

(First step)

$$\max_{\substack{\beta_2 \in \mathcal{B}_2 \\ (z_{12},\ldots,z_{n2}) \in \{0,1\}^n}} \sum_{i=1}^{n} m_{i2}^{B} z_{i2}$$
$$\text{s.t. } \frac{(1, H'_{i2})\beta_2}{C_{i2}} < z_{i2} \leq 1 + \frac{(1, H'_{i2})\beta_2}{C_{i2}} \text{ for } i = 1, \ldots, n,$$

where C_{i2} are constants that should satisfy $C_{i2} > \sup_{\beta_2 \in \mathcal{B}_2} |(1, H'_{i2})\beta_2|$.

Subsequently, the optimization problem in the second step is equivalent to the following MILP problem:

(Second step)

$$\max_{\substack{\beta_1 \in \mathcal{B}_1 \\ (z_{11},\ldots,z_{n1}) \in \{0,1\}^n}} \sum_{i=1}^{n} m_{i1}^{B} z_{i1}$$
$$\text{s.t. } \frac{(1, H'_{i1})\beta_1}{C_{i1}} < z_{i1} \leq 1 + \frac{(1, H'_{i1})\beta_1}{C_{i1}} \text{ for } i = 1, \ldots, n,$$

where C_{i1} are constants that should satisfy $C_{i1} > \sup_{\beta_1 \in \mathcal{B}_1} |(1, H'_{i1})\beta_1|$.

When we specify the dynamic treatment choice problem as the start (stop) time decision problem discussed in Section 2.2, the linear constraints $z_{i2} \geq D_{i1}$ and $D_{i2} \geq z_{i1}$ ($z_{i2} \leq D_{i1}$ and $D_{i2} \leq z_{i1}$) should be added into the MILP problems for the first and second steps, respectively. When we specify the problem as the one-shot treatment decision problem discussed in Section 2.2, the linear constraints $z_{i2} + D_{i1} \leq 1$ and $D_{i2} + z_{i1} \leq 1$ should be added into the MILP problems for the first and second steps, respectively.

H.2 Simultaneous DEWM

In the case of $T = 2$, the optimization problem of the simultaneous DEWM method is equivalent to

$$\max_{(g_1,g_2)\in\mathcal{G}} \sum_{i=1}^{n} \left[m_{i1}^S g_1 + m_{i2}^S g_2 + m_{i3}^S g_1 g_2 \right],$$

where m_{is}^S for $s = 1, 2, 3$ are defined as

$$\begin{aligned}
m_{i1}^S &= \frac{D_{i1}}{e_1(1, H_{i1})} \left(\gamma_1 Y_{i1} + \frac{(1 - D_{i2})\gamma_2 Y_{i2}}{e_2(0, H_{i2})} \right) - \frac{1 - D_{i1}}{e_1(0, H_{i1})} \left(\gamma_1 Y_{i1} + \frac{(1 - D_{i2})\gamma_2 Y_{i2}}{e_2(0, H_{i2})} \right), \\
m_{i2}^S &= \left(\frac{(1 - D_{i1}) D_{i2}}{e_1(0, H_{i1}) e_2(1, H_{i2})} - \frac{(1 - D_{i1})(1 - D_{i2})}{e_1(0, H_{i1}) e_2(0, H_{i2})} \right) \gamma_2 Y_{i2}, \\
m_{i3}^S &= \sum_{(d_1,d_2)\in\{0,1\}^2} \frac{1\{D_{i1} = d_1, D_{i2} = d_2\}\gamma_2 Y_{i2}}{e_1(d_1, H_{i1}) e_2(d_2, H_{i2})}.
\end{aligned}$$

When the class of DTRs is constrained to the class of linear eligibility rules, the above optimization problem is equivalent to the following MILP problem:

$$\begin{aligned}
\max_{\substack{(\beta_1,\beta_2)\in\mathcal{B}_1\times\mathcal{B}_2 \\ (z_{1t},\ldots,z_{nt})_{t=1}^3\in\{0,1\}^{3n}}} & \sum_{i=1}^{n} \left[m_{i1}^S z_{i1} + m_{i2}^S z_{i2} + m_{i3}^S z_{i3} \right] \\
\text{s.t. } & \frac{(1, H_{it}')\beta_t}{C_{it}} < z_{it} \leq 1 + \frac{(1, H_{it}')\beta_t}{C_{it}} \text{ for } i = 1, \ldots, n \text{ and } t = 1, 2, \\
& z_{i3} = z_{i1} z_{i2} \text{ for } i = 1, \ldots, n,
\end{aligned}$$

where C_{it} are constants that should satisfy $C_{it} > \sup_{\beta_t\in\mathcal{B}_t} |(1, H_{it}')\beta_t|$.

When we specify the dynamic treatment choice problem as the start (stop) time decision problem discussed in Section 2.2, the linear constraints $z_{i2} \geq z_{i1}$ ($z_{i2} \leq z_{i1}$) should be added into the MILP problem. When we specify the problem as the one-shot treatment decision problem discussed in Section 2.2, the linear constraint $z_{i1} + z_{i2} \leq 1$ should be added into the MILP problem.

H.3 Budget/Capacity Constraint

The budget/capacity constraints studied in Section 4 can be incorporated into the MILP problem for the simultaneous DEWM. The optimization problem (12) with the class of

linear eligibility score rules is formulated as the following MILP problem:

$$\begin{aligned}
\max_{\substack{(\beta_1,\beta_2)\in\mathcal{B}_1\times\mathcal{B}_2 \\ (z_{1t},\ldots,z_{nt})_{t=1}^{3}\in\{0,1\}^{3n}}} \quad & \sum_{i=1}^{n}\left[m_{i1}^{S}z_{i1}+m_{i2}^{S}z_{i2}+m_{i3}^{S}z_{i3}\right] \\
\text{s.t.} \quad & \frac{(1,H_{it}')\beta_t}{C_{it}} < z_{it} \le 1+\frac{(1,H_{it}')\beta_t}{C_{it}} \text{ for } i=1,\ldots,n \text{ and } t=1,2, \\
& z_{i3}=z_{i1}z_{i2} \text{ for } i=1,\ldots,n, \\
& \frac{1}{n}\sum_{t=1}^{2}\sum_{i=1}^{n}K_{tb}z_{it} \le C_b+\alpha_n \text{ for } b=1,\ldots,B \text{ and } t=1,2,
\end{aligned}$$

where C_{it} are constants that should satisfy $C_{it} > \sup_{\beta_t\in\mathcal{B}_t}|(1,H_{it}')\beta_t|$. The linear constraints in the last line correspond to the budget/capacity constraints.

I Cost and Benefit of Teacher Aide

This appendix outlines the calculation of the cost associated with a full-time teacher aide in our empirical application (Section 7) and evaluates the monetary benefits of the estimated DTRs. Since the outcome measure in our study 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). In line with Krueger (1999), we apply a 3% annual discount rate, discounting all monetary benefits and cost streams back to age 6, the point of intervention.

We begin by assessing first-grade mathematics test scores in terms of the present value of lifetime income, following the discussion in Krueger (1999, Section IV). Estimates from the High School and Beyond sample in Murnane et al. (1995) suggest that male high school seniors who scored one standard deviation higher on the basic math achievement test in 1980 earned 7.93% more six years later, while the corresponding figure for females was 10.98%. The average earnings in the United States in 1996 were $34,705 for men and $20,570 for women for workers aged 18 and older (U.S. Census Bureau, 2006). Assuming constant real earnings and that Murnane et al.'s (1995) estimates apply to the first-grade test scores in the STAR experiment, the present value of the average earnings gain from raising test scores by one point (0.025 standard deviations) is $1,125.26 for men and $923.53 for women, assuming that students enter the workforce at age 20 and retire at

65, and applying the 3% annual discount rate.[4]

Next, we assess the cost of employing a full-time teacher aide, drawing on Word et al. (1990) and Krueger (1999). According to Word et al. (1990), adding a full-time aide in Grades K-3 across Tennessee cost approximately 75 million dollars annually, while reducing class sizes by one-third cost around 196-205 million dollars per year. Thus, the cost of a full-time teacher aide was at most 38% of the cost of class-size reduction. Krueger (1999) estimated that reducing class sizes by one-third would increase per-student costs by roughly $2,151 per year (including capital costs). Based on this, we estimate the cost of a teacher aide to be $817.40 ($= 2,151 \times 0.38$) per student per grade, assuming that kindergarten and grade 1 have equal costs. Applying a 3% real discount rate, the present values of the costs of a full-time teacher aide for grades K and 1 are approximately $817.40 and $793.60, respectively.

Using these estimates, we convert the cost of a full-time teacher aide into test score points. For kindergarten, the cost of a full-time aide corresponds to 0.726 points ($\approx$ 817.4/1125.26) for each male student and 0.885 points ($\approx$ 817.4/923.53) for each female student on the mathematics test. In grade 1, the cost equates to 0.705 points ($\approx$ 793.6/1125.26) for males and 0.859 points ($\approx$ 793.6/923.53) for females. Assuming an equal ratio of male and female students, we estimate the average cost of a full-time aide per student to be equivalent to 0.782 points in kindergarten and 0.806 points in grade 1. These cost values are used in our empirical analysis in Section Section 7. Based on these calculations, we estimate that the DTR $\hat{g}^S$ presented in Table 2 would increase the present value of each student's lifetime earnings, minus the cost of the teacher aide, by an average of $7,428.90.

We are aware that this cost/benefit calculation relies on many assumptions, any of which could prove incorrect. These include: real earnings may increase or decrease over time; the impact of test scores on future earnings may differ from what is assumed; and general equilibrium effects may arise, such as an increase in the overall education level of the population expanding the supply of skilled labor, potentially lowering wage rates for more educated individuals.

[4]Chetty et al. (2011) also estimate the average lifetime earnings gain associated with test score improvements. However, since their analysis focuses on the sum of reading and mathematics test scores rather than mathematics scores alone, we use the results from Murnane et al. (1995) to estimate the average earnings gain from increasing mathematics test scores.

References

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

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

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.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," .

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126, 1593–1660.

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.

Ertefaie, A., J. R. McKay, D. Oslin, and R. L. Strawderman (2021): "Robust Q-learning," *Journal of the American Statistical Association*, 116, 368–381.

Giné, E. and R. Nickl (2016): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, New York: Cambridge University Press.

Jiang, N. and L. Li (2016): "Doubly robust off-policy value evaluation for reinforcement learning," in *International Conference on Machine Learning*, PMLR, 652–661.

Kitagawa, T. and A. Tetenov (2018): "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.

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

Le, H., C. Voloshin, and Y. Yue (2019): "Batch policy learning under constraints," in *International Conference on Machine Learning*, PMLR, 3703–3712.

Massart, P., É. Nédélec, et al. (2006): "Risk bounds for statistical learning," *Annals of Statistics*, 34, 2326–2366.

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

Munos, R. and C. Szepesvári (2008): "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research*, 9.

Murnane, R. J., J. B. Willett, and F. Levy (1995): "The Growing Importance of Cognitive Skills in Wage Determination," *The Review of Economics and Statistics*, 77, 251–266.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89, 846–866.

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

Thomas, P. and E. Brunskill (2016): "Data-efficient off-policy policy evaluation for reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2139–2148.

U.S. Census Bureau (2006): *Historical Income Tables*, (Washington, DC).

Wallace, M. P. and E. E. Moodie (2015): "Doubly-robust dynamic treatment regimen estimation via weighted least squares," *Biometrics*, 71, 636–644.

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

Word, E., J. Johnston, H. P. Bain, B. D. Fulton, J. B. Zaharies, M. N. Lintz, C. M. Achilles, J. Folger, and C. Breda (1990): "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report (1985-1990)." Tennessee State Department of Education.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013): "Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions," *Biometrika*, 100, 681–694.

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