Subsampling-based Methods for Inference of the Mean Outcome under Optimal Treatment Regimes

Abstract

Precision medicine is an emerging medical approach that allows physicians to select the treatment options based on individual patient information. The goal of precision medicine is to identify the optimal treatment regime (OTR) that yields the most favorable clinical outcome on average. Although considerable research has been devoted to estimating the optimal treatment regime (OTR) in the literature, less attention has been paid to statistical inference of the OTR. In this paper, we develop a novel inference method for the mean outcome under an OTR (the optimal value function) based on subsample aggregating (subagging) and refitted cross-validation. The proposed method can be applied to multi-stage studies where treatments are sequentially assigned over time.

Bootstrap aggregating (bagging) and subagging have been recognized as effective variance reduction techniques in hard decision problems (Bühlmann and Yu, 2002). However, it remains unknown whether these approaches can yield valid inference results. We show the proposed confidence interval (CI) for the optimal value function achieves nominal coverage even at nonregular cases where the OTR is not uniquely defined. In addition, due to the variance reduction effect of subagging, our method enjoys certain statistical optimality. Specifically, we prove the squared length of the proposed CI is on average shorter than the CI constructed based on the online one-step method (Luedtke and van der Laan, 2016). Moreover, under certain conditions on the propensity score function, we show the proposed CI is asymptotically narrower than the CI of the "oracle" method which works as well as if the OTR were known. Numerical studies are conducted to back up our theoretical findings.

Keywords: Optimal (dynamic) treatment regime; Optimal value function; Precision medicine; Refitted cross-validation; Subsample aggregating.

1 Introduction

Precision medicine is an emerging medical approach that allows physicians to select the treatment options based on individual patient information. In contrast to the "one-size-fits-all" approach, precision medicine proposes to identify the optimal treatment regime (OTR) that yields the most favorable clinical outcome on average. A treatment regime is a function that maps a patient's baseline covariates to the space of available treatment options. For chronic diseases such as cancer and diabetes, treatment of patients involves a series of decisions. In these applications, it is of considerable interest to estimate the optimal dynamic treatment regime (ODTR) that consists of a list of decision rules for assigning treatment based on a patient's covariates and treatment history.

In the literature, considerable research has been devoted to estimating the OTR (or ODTR). Some popular methods include Q-learning (Watkins and Dayan, 1992; Chakraborty et al., 2010), A-learning (Robins et al., 2000; Murphy, 2003), policy search methods (Zhang et al., 2012, 2013), outcome weighted learning (Zhao et al., 2012, 2015), concordance-assisted learning (Fan et al., 2017; Liang et al., 2017) and decision list-based methods (Zhang et al., 2015, 2016). Prior to adopting any OTR in clinical practice, it is crucial to know the impact of implementing such a policy. This requires to evaluate the mean outcome in the population under an OTR, i.e, the optimal value function. The inference of the optimal value function helps us to evaluate whether the OTR can lead to a clinically meaningful increment value compared to fixed treatment regimes.

Despite the popularity of estimating the OTR, less attention has been paid to statistical inference of the optimal value function. This is an extremely challenging task in the non-regular cases where there is a positive probability that the interaction between treatment and covariates (i.e, the contrast function) is equal to zero. The main challenge lies in that the OTR is unknown and needs to be estimated from the data. Consider the following naive method that first estimates the OTR and then evaluates its mean outcome based on the augmented inverse propensity-score weighted estimator (AIPWE) for the value function (Zhang et al., 2012, 2013). The validity of such a procedure relies on the estimated treatment regime being consistent to the OTR. However, this condition is typically violated

in the non-regular cases (see Section 4.1 in Luedtke and van der Laan, 2016).

Chakraborty et al. (2014) considered inference for the value of an estimated OTR using the m-out-of-n bootstrap. The CI based on this method is valid in the nonregular cases when m grows to infinity at a rate slower than n. However, the length of the CI shrinks at a rate of $m^{-1/2}$. As a result, such CI will be much wider than the CI of our proposed procedure which shrinks at a rate of $n^{-1/2}$. Luedtke and van der Laan (2016) proposed an online one-step estimator that is $n^{-1/2}$-consistent to the optimal value function. Their method mimics the online prediction algorithms and recursively updates the initial estimated OTR and value function using new observations. Based on the online one-step estimator, they developed a valid inference procedure. However, their procedure relies on a data ordering. The online one-step estimator can be sensitive to the order of the data, especially when the sample size is small.

In this paper, we develop a novel inference method for the optimal value function based on subsample aggregating (subagging) and refitted cross-validation. Specifically, we estimate the OTR based on a random subsample of the data and evaluate its value based on the remaining data using AIPWE. We then iterate this procedure multiple times. Our final estimator is defined as an average of all value estimators. To compute AIPWE, we need to estimate the propensity score function and the conditional mean of the response given treatment and predictors. To avoid model misspecification and gain efficiency, we propose to estimate these function nonparametrically and use a sample-splitting method to construct AIPWE. The use of sample-splitting helps reduce the bias of AIPWE resulting from the biases of the estimated propensity score and conditional mean functions.

Bootstrap aggregating (bagging) and subagging have been recognized as effective variance reduction techniques in hard decision problems (Bühlmann and Yu, 2002). However, it remains unknown whether these procedures can yield valid inference results. We show the proposed estimator is asymptotically normal even at nonregular cases. We further provide a consistent estimator for its asymptotic variance and derive a Wald-type CI for the optimal value function. Due to the variance reduction effect of subagging, our method enjoys certain statistical optimality. More specifically, we prove that the squared length of

the proposed CI is on average shorter than the CI constructed based on the online one-step method (Luedtke and van der Laan, 2016) in the nonregular cases. In addition, under certain conditions on the propensity score function, our proposed CI is asymptotically narrower than the CI of the "oracle" method which works as well as if the OTR were known. These theoretical findings are further supported by extensive numerical studies.

Moreover, the proposed method can be applied to multi-stage studies to evaluate the mean outcome under an ODTR. In their supplementary material, Luedtke and van der Laan (2016) extended their method to a two-stage study and proved the validity of their inference procedure. However, some of the technical conditions they imposed are not interpretable. For example, they assumed that their estimator has bias of the order $o_p(n^{-1/2})$ in (A.C5) but didn't give sufficient conditions under which (A.C5) holds. In this paper, we provide more easily interpretable conditions, and show that our proposed value estimator is asymptotically unbiased and our CI achieves nominal coverage, as long as the estimated contrast function at each stage satisfies certain convergence rates and the true contrast function at each stage satisfies certain margin conditions. Besides, we allow the number of treatment stages to be an arbitrary fixed integer.

The rest of the paper is organized as follows. In Section 2, we introduce our inference procedure in a point treatment study. In Section 3, we discuss the asymptotic optimality of our proposed method. Section 4 contains the extension to multi-stage studies. Simulation studies are conducted in Section 5. In Section 6, we apply the proposed method to a real dataset. The proof of Theorem 2.1 is given in the Appendix. Other proofs are provided in the supplementary article.

2 Point treatment study

2.1 Optimal treatment regime

We begin by considering a single stage study with two treatments. Let $\boldsymbol{X}_0 \in \mathbb{X}$ be a patient's baseline covariates, $A_0 \in \{0, 1\}$ denote the treatment a patient receives, and Y_0 denote a patient's clinical outcome (the larger the better by convention). A treatment

regime $d(\cdot)$ is a deterministic function that maps $\mathbb{X}$ to $\{0,1\}$. Let $Y_0^*(0)$ and $Y_0^*(1)$ be a patient's potential outcomes, representing the response he/she would get if treated by treatment 0 and 1, respectively. In addition, define the potential outcome

$$Y_0^*(d) = Y_0^*(0)\{1 - d(\boldsymbol{X}_0)\} + Y_0^*(1)d(\boldsymbol{X}_0),$$

representing the response a patient would have if treated according to a treatment regime d. Let $V(d) = \mathrm{E}\{Y_0^*(d)\}$. An OTR d^{opt} is defined as the maximizer of the expected potential outcome $V(d)$ among the set of all possible treatment regimes, i.e,

$$d^{opt} \equiv \arg\max_d V(d).$$

However, the OTR may not be unique. Let $\mathcal{D}^{opt}$ denote the set of all OTRs, i.e,

$$\mathcal{D}^{opt} = \{d_0 : V(d_0) = \max_d V(d)\}.$$

Assume the following two assumptions hold.

(A1.) SUTVA: $Y_0 = (1 - A_0)Y_0^*(0) + A_0 Y_0^*(1)$.

(A2.) No unmeasured confounders: $Y_0^*(0), Y_0^*(1) \perp\!\!\!\perp A_0 | \boldsymbol{X}_0$.

Define the contrast function

$$\tau(\boldsymbol{x}) \equiv \mathrm{E}\{Y_0 | A_0 = 1, \boldsymbol{X}_0 = \boldsymbol{x}\} - \mathrm{E}\{Y_0 | A_0 = 0, \boldsymbol{X}_0 = \boldsymbol{x}\}.$$

The following lemma relates OTR to the function $\tau(\cdot)$.

Lemma 2.1. *Let* $\mathbb{X}_1 = \{\boldsymbol{x} \in \mathbb{X} : \tau(\boldsymbol{x}) > 0\}$ *and* $\mathbb{X}_2 = \{\boldsymbol{x} \in \mathbb{X} : \tau(\boldsymbol{x}) < 0\}$. *Assume (A1), (A2) hold, and* $E|\tau(\boldsymbol{X}_0)| < \infty$. *Then, for any* $d \in \mathcal{D}^{opt}$, *we have*

$$Pr(\boldsymbol{X}_0 \in \mathbb{X}_1 \cap \mathbb{X}_{2,d}) = 0 \quad and \quad Pr(\boldsymbol{X}_0 \in \mathbb{X}_2 \cap \mathbb{X}_{1,d}) = 0, \tag{1}$$

where $\mathbb{X}_{1,d} = \{\boldsymbol{x} \in \mathbb{X} : d(\boldsymbol{x}) = 1\}$ *and* $\mathbb{X}_{2,d} = \{\boldsymbol{x} \in \mathbb{X} : d(\boldsymbol{x}) = 0\}$. *Conversely, for any treatment regime* d *satisfying* (1), *we have* $d \in \mathcal{D}^{opt}$.

Lemma 2.1 implies that $d^{opt,0} \in \mathcal{D}^{opt}$ where

$$d^{opt,0}(\boldsymbol{x}) = \mathbb{I}\{\tau(\boldsymbol{x}) > 0\}, \quad \forall \boldsymbol{x} \in \mathbb{X}, \tag{2}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function. Let $V_0 = \max_d V(d) = V(d^{opt,0})$. Our objective is to construct confidence intervals (CIs) for V_0.

2.2 Sample-split estimation and subsample aggregation

For $a = 0, 1$ and any $\boldsymbol{x} \in \mathbb{X}$, define the propensity score function $\pi(a, \boldsymbol{x}) = \Pr(A_0 = a|\boldsymbol{X}_0 = \boldsymbol{x})$ and the conditional mean function $h(a, \boldsymbol{x}) = \mathrm{E}(Y_0|A_0 = a, \boldsymbol{X}_0 = \boldsymbol{x})$. The observed data can be summarized as $\{O_i = (\boldsymbol{X}_i, A_i, Y_i), i = 1, \ldots, n\}$ where O_i's are i.i.d copies of $O_0 = (\boldsymbol{X}_0, A_0, Y_0)$. Let $\pi^*(\cdot, \cdot)$ and $h^*(\cdot, \cdot)$ denote some estimators for $\pi(\cdot, \cdot)$ and $h(\cdot, \cdot)$. For any $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ and any treatment regime $d(\cdot)$, consider the following augmented inverse propensity-score weighted estimator (AIPWE, Zhang et al., 2012) for $V(d)$,

$$\widehat{V}_{\mathcal{I}}(d; \pi^*, h^*) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(\frac{\mathrm{g}\{A_i, d(\boldsymbol{X}_i)\}}{\pi^*(A_i, \boldsymbol{X}_i)} \{Y_i - h^*(A_i, \boldsymbol{X}_i)\} + h^*\{d(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right),$$

where $|\mathcal{I}|$ denotes the number of elements in $\mathcal{I}$, $\mathrm{g}(y, z) = yz + (1-y)(1-z)$ for all $y, z \in \mathbb{R}$, and π_i is a shorthand for $\pi(\boldsymbol{X}_i)$. The above estimator is consistent when either π^* or h^* is consistent. To better illustrate our method, in the following, we assume functions π and h are known. In Section 2.3, we allow these functions to be estimated from the observed dataset. We use a shorthand and write $\widehat{V}_{\mathcal{I}}(d) = \widehat{V}_{\mathcal{I}}(d; \pi, h)$ for any $\mathcal{I}$ and $d(\cdot)$.

To estimate the optimal value function, we need to estimate the OTR first. We consider the class of plug-in classifiers. More specifically, let $\widehat{\tau}_{\mathcal{I}}(\cdot)$ denote the estimated contrast function based on the sub-dataset $\{O_i\}_{i \in \mathcal{I}}$ and define $\widehat{d}_{\mathcal{I}}(\cdot) = \mathbb{I}\{\widehat{\tau}_{\mathcal{I}}(\cdot) > 0\}$. We may set $\widehat{\tau}_{\mathcal{I}}(\cdot) \equiv 0$ if either $\sum_{i \in \mathcal{I}} A_i = 0$ or $\sum_{i \in \mathcal{I}} \{1 - A_i\} = 0$. Let $\mathcal{I}_0 = \{1, 2, \ldots, n\}$. To obtain valid CI for the optimal value function, we apply sample-split estimation with subagging.

More specifically, we estimate V_0 by

$$\widehat{V}_\infty^* = \frac{1}{\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \widehat{V}_{\mathcal{I}^c}(\widehat{d}_{\mathcal{I}}),$$

where s_n is some diverging sequence, and $\mathcal{I}^c$ denotes the complement of $\mathcal{I}$.

For any integer $j > 0$, define $p_j(\boldsymbol{x}) = \Pr(\widehat{\tau}_{\{1,2,\ldots,j\}}(\boldsymbol{x}) > 0)$. For any $\mathcal{I} \subseteq \mathcal{I}_0$ with $|\mathcal{I}| = s_n$, let $R_{\mathcal{I}}(\boldsymbol{x}) = \widehat{d}_{\mathcal{I}}(\boldsymbol{x}) - p_{s_n}(\boldsymbol{x})$. It is immediate to see that $\mathrm{E}R_{\mathcal{I}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{X}$. For any $y \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{X}$, let $h(y, \boldsymbol{x}) = yh(1, \boldsymbol{x}) + (1-y)h(0, \boldsymbol{x})$. By definition, we have

$$\begin{aligned}
\widehat{V}_\infty^* &= \frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \left(\frac{\mathrm{g}\{A_i, \widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h\{\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right) \\
&= \underbrace{\frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \left(\frac{\mathrm{g}\{A_i, p_{s_n}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h\{p_{s_n}(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right)}_{\eta_1} \\
&+ \underbrace{\frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \left(\frac{(2A_i-1)R_{\mathcal{I}}(\boldsymbol{X}_i)}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + \tau(\boldsymbol{X}_i) R_{\mathcal{I}}(\boldsymbol{X}_i) \right)}_{\eta_2}.
\end{aligned}$$

When $s_n = o(n)$, we can show $\eta_2 = o_p(n^{-1/2})$. Notice that $p_{s_n}(\cdot)$ is a deterministic function of $\boldsymbol{x}$. Let $\mathcal{I}_{(-i)} = \mathcal{I}_0 - \{i\}$, we have

$$\begin{aligned}
\eta_1 &= \frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{i=1}^n \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_{(-i)} \\ |\mathcal{I}|=s_n}} \left(\frac{\mathrm{g}\{A_i, p_{s_n}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h\{p_{s_n}(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right) \\
&= \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathrm{g}\{A_i, p_{s_n}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h\{p_{s_n}(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right).
\end{aligned}$$

As a result, $\sqrt{n}\widehat{V}_\infty^*$ is asymptotically equivalent to a sum of i.i.d mean zero random variables. Below, we formally establish these results. We need the following conditions.

(A3) Assume there exists some positive constant c_0 such that $\inf_{\boldsymbol{x}\in\mathbb{X},a=0,1} \pi(a, \boldsymbol{x}) \geq c_0$.

(A4) Assume $\sup_{\boldsymbol{x}\in\mathbb{X},a=0,1} \mathrm{E}[\{Y_0^*(a)\}^2 | \boldsymbol{X}_0 = \boldsymbol{x}] = O(1)$.

(A5) Assume there exist some positive constants $\bar{c}$, α and δ_0 such that

$$\Pr(0 < |\tau(\boldsymbol{X}_0)| < t) \leq \bar{c}t^{\alpha}, \qquad \forall 0 < t \leq \delta_0.$$

(A6) Assume there exists some constant $\kappa_0 > (\alpha+2)/(2\alpha+2)$ such that

$$\mathrm{E}|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_0) - \tau(\boldsymbol{X}_0)|^2 = O(|\mathcal{I}|^{-\kappa_0}),$$

for any $\mathcal{I} \subseteq \mathcal{I}_0$.

Condition (A3) holds when $\pi(0,\cdot)$ and $\pi(1,\cdot)$ are positive constants, as in randomized studies. Condition (A4) automatically holds when the potential outcomes $Y_0^*(0)$ and $Y_0^*(1)$ are bounded. Condition (A5) is very similar to the margin assumption from Audibert and Tsybakov (2007). It holds with $\alpha = 1$ when $\tau(\boldsymbol{X}_0)$ has a bounded probability density function near 0. This condition was also considered by Qian and Murphy (2011) and Luedtke and Chambaz (2017) to establish the convergence rates of the value function under an estimated OTR. In (A6), we assume the estimated contrast function shall satisfy certain convergence rates. These rates are available for most often used nonparametric approaches including spline methods (Zhou et al., 1998), kernel ridge regression (Steinwart and Christmann, 2008; Zhang et al., 2013), tree-based methods (Zhu et al., 2015) and random forests (Biau, 2012). When (A5) and (A6) hold, we can show that $V(\widehat{d}_{\mathcal{I}}) = V_0 + o_p(|\mathcal{I}|^{-1/2})$ for any $\mathcal{I} \subseteq \mathcal{I}_0$.

We present our main results below. For any two sequences $\{a_n\}, \{b_n\}$, we write $a_n \asymp b_n$ if there exist some universal constants c, $C > 0$ such that $cb_n \leq a_n \leq Cb_n$.

Theorem 2.1. *Assume (A1)-(A6) hold, and s_n satisfies $s_n \asymp n^{\beta_0}$ for some $(2+\alpha)/\{\kappa_0(2+2\alpha)\} < \beta_0 < 1$. Then, we have*

$$\widehat{V}_{\infty}^* = \eta_1 + o_p(n^{-1/2}) \;\; \textit{and} \;\; V_0 = E\eta_1 + o(n^{-1/2}).$$

Theorem 2.1 implies that $\sqrt{n}(\widehat{V}_\infty^* - V_0) = \sqrt{n}\{\eta_1 - \mathrm{E}\eta_1\} + o_p(1)$. For $j = 1, 2, \ldots$, let

$$\sigma_j^2 = \mathrm{Var}\left(\frac{\mathrm{g}\{A_0, p_j(\boldsymbol{X}_0)\}}{\pi(A_0, \boldsymbol{X}_0)} \{Y_0 - h(A_0, \boldsymbol{X}_0)\} + h\{p_j(\boldsymbol{X}_0), \boldsymbol{X}_0\} \right).$$

By (A3) and (A4), we have $\sup_{j \geq 1} \sigma_j = O(1)$. Assume $\liminf_n \sigma_n > 0$. By central limit theorem, we have

$$\frac{\sqrt{n}(\widehat{V}_\infty^* - V_0)}{\sigma_{s_n}} \xrightarrow{d} N(0, 1).$$

For any $z_1, \ldots, z_n \in \mathbb{R}$, let $\widehat{s.e}^2(\{z_i\}_{i=1}^n)$ denote the sample variance estimator, i.e, $\widehat{s.e}^2(\{z_i\}_{i=1}^n) = \sum_{i=1}^n (z_i - \bar{z})^2/(n-1)$ where $\bar{z} = \sum_{i=1}^n z_i/n$. The asymptotic variance $\sigma_{s_n}^2$ can be consistently estimated by

$$\widehat{\sigma}_\infty^{*2} = \widehat{s.e}^2 \left(\left\{ \frac{\mathrm{g}\{A_i, \widehat{d}_{s_n}^{(-i)}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h\{\widehat{d}_{s_n}^{(-i)}(\boldsymbol{X}_i), \boldsymbol{X}_i\} \right\}_{i=1}^n \right),$$

where

$$\widehat{d}_{s_n}^{(-i)}(\boldsymbol{x}) = \frac{1}{\binom{n-1}{s_n}} \sum_{\substack{\mathcal{I} \subseteq \mathcal{I}_{(-i)} \\ |\mathcal{I}| = s_n}} \widehat{d}_{\mathcal{I}}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{X}.$$

Notice that it is intractable to compute $\widehat{d}_{\mathcal{I}}$ over all possible size s_n subsamples of the training data. In practice, we can estimate $\widehat{V}_\infty^*$ based on Monte Carlo approximations. More specifically, for a sufficiently large integer B, set

$$\widehat{V}_B^* = \frac{1}{B} \sum_{b=1}^B \widehat{V}_{\mathcal{I}_b^c}(\widehat{d}_{\mathcal{I}_b}), \tag{3}$$

where the subsets $\mathcal{I}_1, \ldots, \mathcal{I}_B$ are drawn uniformly from the set

$$\mathcal{S}_{N_0, s_n} = \left\{ \mathcal{I} \subseteq \mathcal{I}_0 : |\mathcal{I}| = s_n, N_0 \leq \sum_{i \in \mathcal{I}} A_i \leq s_n - N_0 \right\},$$

for some positive integer N_0. Here, the constraints $N_0 \leq \sum_{i\in\mathcal{I}} A_i \leq s_n - N_0$ guarantee that the function $\tau(\cdot)$ is estimable based on the sub-dataset $\{O_i\}_{i\in\mathcal{I}}$. Define

$$\widehat{\sigma}_B^{*2} = \widehat{s.e}^2 \left(\left\{ \frac{\mathrm{g}\{A_i, \widehat{d}^{(-i)}_{s_n,B}(\boldsymbol{X}_i)\}}{\pi(A_i, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h(\widehat{d}^{(-i)}_{s_n,B}(\boldsymbol{X}_i), \boldsymbol{X}_i) \right\}_{i=1}^{n} \right), \tag{4}$$

where

$$\widehat{d}^{(-i)}_{s_n,B}(\boldsymbol{X}_i) = \frac{1}{n^{(i)}} \sum_{b:\{i\notin\mathcal{I}_b\}} \widehat{d}_{\mathcal{I}_b}(\boldsymbol{X}_i),$$

and $n^{(i)} = \sum_{b=1}^{B} \mathbb{I}(i \notin \mathcal{I}_b)$. The corresponding two-sided CI for V_0 is given by

$$\left[\widehat{V}_B^* - \frac{2z_{\alpha/2}\widehat{\sigma}_B^*}{\sqrt{n}}, \widehat{V}_B^* + \frac{2z_{\alpha/2}\widehat{\sigma}_B^*}{\sqrt{n}} \right].$$

2.3 Unknown propensity score and conditional mean functions

In practice, the conditional mean function $h(\cdot,\cdot)$ is unknown to us. In observational studies, the propensity score function $\pi(\cdot,\cdot)$ also needs to be estimated from the data. We propose to estimate these function via some nonparametric methods. For any $\mathcal{I} \subseteq \mathcal{I}_0$, denoted by $\widehat{h}_{\mathcal{I}}$ and $\widehat{\pi}_{\mathcal{I}}$ the corresponding estimators for h and π, based on the sub-dataset $\{O_i\}_{i\in\mathcal{I}}$. For simplicity, assume $n - s_n = 2t_n$ for some integer $t_n > 0$. We detail our procedure in the following algorithm.

Step 1. Input observations $\{O_i\}_{i=1,\ldots,n}$, $0 < \alpha < 1$, and integers s_n, N_0 and B.

Step 2. For $b = 1, \ldots, B$,

(i) Draw a subset $\mathcal{I}_b$ from $\mathcal{S}_{N_0,s_n}$ uniformly at random.

(ii) Randomly partition $\mathcal{I}_b^c$ into 2 disjoint subsets $\mathcal{I}_b^{c(1)}$ and $\mathcal{I}_b^{c(2)}$ of equal sizes t_n.

(iii) For $j = 1, 2$, let $\mathcal{I}_b^{(j)} = \mathcal{I}_b \cup \mathcal{I}_b^{c(j)}$. Obtain the estimators $\widehat{d}_{\mathcal{I}_b}$, $\widehat{\pi}_{\mathcal{I}_b^{(1)}}$, $\widehat{\pi}_{\mathcal{I}_b^{(2)}}$, $\widehat{h}_{\mathcal{I}_b^{(1)}}$ and $\widehat{h}_{\mathcal{I}_b^{(2)}}$.

Step 3. Compute

$$\widehat{V}_B = \frac{1}{2B} \sum_{b=1}^{B} \left(\widehat{V}_{\mathcal{I}_b^{c(2)}}(\widehat{d}_{\mathcal{I}_b}; \widehat{\pi}_{\mathcal{I}_b^{(1)}}, \widehat{h}_{\mathcal{I}_b^{(1)}}) + \widehat{V}_{\mathcal{I}_b^{c(1)}}(\widehat{d}_{\mathcal{I}_b}; \widehat{\pi}_{\mathcal{I}_b^{(2)}}, \widehat{h}_{\mathcal{I}_b^{(2)}}) \right),$$

and $\widehat{\sigma}_B^2 = \widehat{s.e}^2(\{\widehat{V}^{(i)}\}_{i=1}^n)$ where

$$\widehat{V}^{(i)} = \frac{1}{n^{(i)}} \sum_{b=1}^{B} \left(\widehat{V}_{\{i\}}(\widehat{d}_{\mathcal{I}_b}; \widehat{\pi}_{\mathcal{I}_b^{(1)}}, \widehat{h}_{\mathcal{I}_b^{(1)}}) \mathbb{I}(i \notin \mathcal{I}_b^{(1)}) + \widehat{V}_{\{i\}}(\widehat{d}_{\mathcal{I}_b}; \widehat{\pi}_{\mathcal{I}_b^{(2)}}, \widehat{h}_{\mathcal{I}_b^{(2)}}) \mathbb{I}(i \notin \mathcal{I}_b^{(2)}) \right),$$

and $n^{(i)} = \sum_{b=1}^{B} \mathbb{I}(i \notin \mathcal{I}_b)$.

Step 4. Output

$$\left[\widehat{V}_B - \frac{z_{\alpha/2}\widehat{\sigma}_B}{\sqrt{n}}, \widehat{V}_B + \frac{z_{\alpha/2}\widehat{\sigma}_B}{\sqrt{n}} \right]. \tag{5}$$

Notice that we apply a two-fold cross-validation procedure in Step 2 and 3. More generally, one can use K-fold cross-validation to construct the estimators $\widehat{V}_B$ and $\widehat{\sigma}_B^2$, for any fixed integer $K \geq 2$. The following theorem proves the validity of the CI in (5).

Theorem 2.2. *Assume $B \gg n$, $\liminf_n \sigma_n > 0$,*

$$Pr\left(\inf_{\mathcal{I} \in \mathcal{I}_0, \boldsymbol{x} \in \mathbb{X}, a=0,1} \widehat{\pi}_{\mathcal{I}}(a, \boldsymbol{x}) \geq c^* \right) = 1, \tag{6}$$

for some constant $c^* > 0$. In addition, assume

$$\max_{a=0,1} E|\widehat{\pi}_{\mathcal{I}}(a, \boldsymbol{X}_0) - \pi(a, \boldsymbol{X}_0)|^2 = o(|\mathcal{I}|^{-1/2}), \max_{a=0,1} E|\widehat{h}_{\mathcal{I}}(a, \boldsymbol{X}_0) - h(a, \boldsymbol{X}_0)|^2 = o(|\mathcal{I}|^{-1/2}), \tag{7}$$

for any $\mathcal{I} \subseteq \mathcal{I}_0$. Then, under the conditions in Theorem 2.1, we have

$$\frac{\sqrt{n}(\widehat{V}_B - V_0)}{\widehat{\sigma}_B} \xrightarrow{d} N(0, 1).$$

In Theorem 2.2, we require the estimator $\widehat{\pi}_{\mathcal{I}}$ to be uniformly bounded away from 0. To satisfy this condition, for a given estimated propensity function $\tilde{\pi}$, we can set $\widehat{\pi}$ to be the following truncated estimator:

$$\widehat{\pi}_{\mathcal{I}}(\cdot, \cdot) = \max\left(\tilde{\pi}_{\mathcal{I}}(\cdot, \cdot), \varepsilon_0\right), \qquad \forall \mathcal{I} \subseteq \mathcal{I}_0,$$

for some sufficiently small constant $\varepsilon_0 > 0$. In (7), we require the estimated propensity score and conditional mean functions to satisfy certain convergence rates. These conditions guarantee that $\widehat{V}_B$ and $\widehat{\sigma}_B^2$ are asymptotically equivalent to $\widehat{V}_B^*$ and $\widehat{\sigma}_B^{*2}$, defined in (3) and (4), respectively.

Theorem 2.2 shows the asymptotic normality of $\sqrt{n}(\widehat{V}_B - V_0)/\widehat{\sigma}_B$. As a result, the two-sided CI defined in (5) has asymptotically nominal coverage probabilities. Moreover, it also implies that $\widehat{V}_B - z_\alpha \widehat{\sigma}_B/\sqrt{n}$ is an asymptotic $1-\alpha$ lower confidence bound for V_0.

3 Asymptotic optimality

This section discusses the optimality of the proposed method. The length of the proposed CI (see (5)) is given by $L(\widehat{V}_B, \alpha) = 2z_{\alpha/2}\widehat{\sigma}_B/\sqrt{n}$. Under the given conditions in Theorem 2.2, the estimator $\widehat{\sigma}_B$ is consistent to σ_{s_n} and we can show

$$\sqrt{n}L(\widehat{V}_B, \alpha) = 2z_{\alpha/2}\sigma_{s_n} + o_p(1), \tag{8}$$

and

$$n\mathrm{E}L^2(\widehat{V}_B, \alpha) = 4z_{\alpha/2}^2\sigma_{s_n}^2 + o(1). \tag{9}$$

Before presenting our main results, we introduce the following condition.

(A7) For any $\boldsymbol{x} \in \mathbb{X}$ with $\tau(\boldsymbol{x}) = 0$, assume the following holds:

$$\Pr\left(\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) > 0\right) \to 1/2, \qquad \text{as } |\mathcal{I}| \to \infty.$$

Assume $\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) - \tau(\boldsymbol{x})$ is asymptotically normal, i.e,

$$\frac{\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) - \tau(\boldsymbol{x})}{\sigma^*_{|\mathcal{I}|}} \xrightarrow{d} N(0,1), \tag{10}$$

for some sequence $\sigma_n^* \to 0$. Then it follows from the definition of weak convergence that

$$\Pr\left(\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) > 0\right) = \Pr\left(\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) - \tau(\boldsymbol{x}) > 0\right) = \Pr\left(\frac{\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}) - \tau(\boldsymbol{x})}{\sigma^*_{|\mathcal{I}|}} > 0\right) \to \frac{1}{2}, \quad \forall \boldsymbol{x} \text{ with } \tau(\boldsymbol{x}) = 0.$$

Notice that the condition (10) holds for a wide variety of nonparametric estimators $\widehat{\tau}_{\mathcal{I}}$ computed by kernel smoothing methods (Härdle, 1990), spline methods (Zhou et al., 1998), kernel ridge regression (Zhao et al., 2016), random forests (Wager and Athey, 2017), etc.

For simplicity, throughout this section, we assume the following semiparametric regression model for Y_0:

$$Y_0 = h(0, \boldsymbol{X}_0) + A_0\tau(\boldsymbol{X}_0) + e_0, \tag{11}$$

where e_0 is a mean zero random error term independent of $\boldsymbol{X}_0, A_0$. Let $\sigma_0^2 = \mathrm{Var}(e_0) > 0$.

3.1 Comparison with the online one-step estimator

For $j = 1, \ldots, n$, let $\mathcal{I}_{(j)} = \{1, 2, \ldots, j\}$. Let $\{l_n\}_n$ be a sequence of nonnegative integer with $l_n < n$. The online one-step estimator is defined as

$$\widehat{V}^{on} = \left(\sum_{j=l_n}^{n-1} \widetilde{\sigma}^{-1}_{\mathcal{I}_{(j)}}\right)^{-1} \left(\sum_{j=l_n}^{n-1} \widetilde{\sigma}^{-1}_{\mathcal{I}_{(j)}} \widehat{V}_{\{j+1\}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}})\right),$$

where $\widetilde{\sigma}^2_{\mathcal{I}_{(j)}}$ stands for some consistent estimator of

$$\widetilde{\sigma}^2_0(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}}) = \mathrm{Var}\left(\widehat{V}_{\{j+1\}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}}) \middle| \{O_i\}_{i\in\mathcal{I}_{(j)}}\right),$$

computed based on the observations $\{O_i\}_{i\in\mathcal{I}_{(j)}}$.

Under the conditions in Theorem 2 of Luedtke and van der Laan (2016), it follows from martingale central limit theorem that

$$\frac{\sqrt{n - l_n}(\widehat{V}^{on} - V_0)}{\widehat{\sigma}^{on}} \xrightarrow{d} N(0, 1),$$

where $\widehat{\sigma}^{on} = \{\sum_{j=l_n}^{n-1} \widetilde{\sigma}_{\mathcal{I}_{(j)}}^{-1}/(n-l_n)\}^{-1}$. The corresponding two-sided CI for V_0 is given by

$$\left[\widehat{V}^{on} - z_{\alpha/2}\frac{\widehat{\sigma}^{on}}{\sqrt{n-l_n}}, \widehat{V}^{on} + z_{\alpha/2}\frac{\widehat{\sigma}^{on}}{\sqrt{n-l_n}}\right]. \tag{12}$$

Assume $l_n \to \infty$, under the same conditions as Theorem (5), we can show that

$$\widehat{\sigma}^{on} = \left(\frac{\sum_{j=l_n}^{n-1} \widetilde{\sigma}_0^{-1}(\widehat{d}_{\mathcal{I}_{(j)}}; \pi, h)}{n-l_n}\right)^{-1} + o_p(1). \tag{13}$$

The first term on the RHS of the above expression is a random variable depending on $\{O_i\}_{i\in\mathcal{I}_{(n-1)}}$. In the nonregular cases, the conditional variance $\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(j)}}; \pi, h)$ may not converge to a deterministic quantity. As a result, the length of (12) will fluctuate randomly. Therefore, we focus on comparing the average squared length of (12) with that of our proposed CI.

When $\{\widetilde{\sigma}_{\mathcal{I}_{(j)}}\}_{j=l_n,\ldots,n-1}$ and $\{\widetilde{\sigma}_0(\widehat{d}_{\mathcal{I}_{(j)}}; \pi, h)\}_{j=l_n,\ldots,n-1}$ are uniformly bounded from above, it follows from (13) that

$$\mathrm{E}(\widehat{\sigma}^{on})^2 = \mathrm{E}\left(\frac{\sum_{j=l_n}^{n-1} \widetilde{\sigma}_0^{-1}(\widehat{d}_{\mathcal{I}_{(j)}}; \pi, h)}{n-l_n}\right)^{-2} + o(1).$$

When $l_n = o(n)$, the length of (12) satisfies

$$n\mathrm{E}L^2(\widehat{V}^{on}, \alpha) = 4z_{\alpha/2}^2 \mathrm{E}\left(\frac{\sum_{j=l_n}^{n-1} \widetilde{\sigma}_0^{-1}(\widehat{d}_{\mathcal{I}_{(j)}}; \pi, h)}{n-l_n}\right)^{-2} + o(1). \tag{14}$$

Theorem 3.1. *Assume* (9), (11), (14), *(A1)-(A7) hold,* $s_n, l_n \to \infty$. *Then, we have*

$$nEL^2(\widehat{V}^{on}, \alpha) - nEL^2(\widehat{V}_B, \alpha) \geq \frac{z_{\alpha/2}^2 \sigma_0^2}{2c_0^2} Pr\{\tau(\boldsymbol{X}_0) = 0\} + o(1),$$

where c_0 *is defined in Condition (A3).*

Theorem 3.1 implies that the expected squared length of (12) is asymptotically larger than that of the proposed CI in the nonregular cases. The difference depends on $\Pr\{\tau(\boldsymbol{X}_0) =$

0}, which measures the degree of nonregularity.

In the following, we sketch a few lines to see why our proposed CI achieves smaller squared length on average. By the delta method, we have

$$\mathrm{E}\left(\frac{\sum_{j=l_n}^{n-1}\widetilde{\sigma}_0^{-1}(\widehat{d}_{\mathcal{I}_{(j)}};\pi,h)}{n-l_n}\right)^{-2} \approx \left(\frac{\sum_{j=l_n}^{n-1}\{\mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(j)}};\pi,h)\}^{-1/2}}{n-l_n}\right)^{-2}.$$

Under the given conditions, $\mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(j)}};\pi,h)$ converges to a fixed function as $j\to\infty$. Since $s_n, l_n \to \infty$, we have

$$\begin{aligned}\left(\frac{\sum_{j=l_n}^{n-1}\{\mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(j)}};\pi,h)\}^{-1/2}}{n-l_n}\right)^{-2} &= \left(\frac{\sum_{j=l_n}^{n-1}\{\mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(s_n)}};\pi,h)\}^{-1/2}}{n-l_n}\right)^{-2} + o(1)\\ &= \mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(s_n)}};\pi,h) + o(1).\end{aligned}$$

By definition, we have $\sigma_{s_n}^2 = \widetilde{\sigma}_0^2(p_{s_n};\pi,h)$ and $p_{s_n}(\boldsymbol{x}) = \mathrm{E}\widehat{d}_{\mathcal{I}_{(s_n)}}(\boldsymbol{x})$. The function $\widetilde{\sigma}_0^2(d;\pi,h)$ is convex in d. Therefore, it follows from Jensen's inequality that

$$\mathrm{E}\widetilde{\sigma}_0^2(\widehat{d}_{\mathcal{I}_{(s_n)}};\pi,h) \geq \widetilde{\sigma}_0^2(\mathrm{E}\widehat{d}_{\mathcal{I}_{(s_n)}};\pi,h).$$

This together with (9) and (14) yields that $n\mathrm{E}L^2(\widehat{V}^{on},\alpha) \geq n\mathrm{E}L^2(\widehat{V}_B,\alpha) + o(1)$.

3.2 Beyond oracle property

In this section, we compare the proposed CI with the CI based on the oracle method. The oracle knew the set of optimal treatment regimes $\mathcal{D}^{opt}$ ahead of time. When functions π and h are known, we can estimate V_0 by $\widehat{V}_{\mathcal{I}_0}(d^{opt};\pi,h)$ for an arbitrary $d^{opt}\in\mathcal{D}^{opt}$. To deal with unknown propensity score and conditional mean functions, we can construct our estimator based on the following cross-validation procedure:

Step 1 Input observations $\{O_i\}_{i\in\mathcal{I}_0}$, $0<\alpha<1$.

Step 2 Randomly partition $\mathcal{I}_0$ into 2 disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$ of equal sizes, assuming the sample size n is an even integer.

Step 3 Obtain the estimators $\widehat{\pi}_{\mathcal{I}_j}$ and $\widehat{h}_{\mathcal{I}_j}$ for $j = 1, 2$. For any $d^{opt} \in \mathcal{D}^{opt}$, compute

$$\widehat{V}^{or}(d^{opt}) = \frac{1}{2}\left(\widehat{V}_{\mathcal{I}_1}(d^{opt}; \widehat{\pi}_{\mathcal{I}_2}, \widehat{h}_{\mathcal{I}_2}) + \widehat{V}_{\mathcal{I}_2}(d^{opt}; \widehat{\pi}_{\mathcal{I}_1}, \widehat{h}_{\mathcal{I}_1})\right),$$
$$\widehat{\sigma}^{or}(d^{opt}) = \left\{\frac{1}{n-1}\sum_{j=1}^{2}\sum_{i\in\mathcal{I}_j}\left(\widehat{V}_{\{i\}}(d^{opt}; \widehat{\pi}_{\mathcal{I}_j^c}, \widehat{h}_{\mathcal{I}_j^c}) - \widehat{V}(d^{opt})\right)^2\right\}^{1/2}.$$

Step 4 Output

$$\left[\widehat{V}^{or}(d^{opt}) - \frac{z_{\alpha/2}\widehat{\sigma}^{or}(d^{opt})}{\sqrt{n}}, \widehat{V}^{or}(d^{opt}) + \frac{z_{\alpha/2}\widehat{\sigma}^{or}(d^{opt})}{\sqrt{n}}\right]. \tag{15}$$

The CI in (15) is valid. Under conditions (6) and (7), we can show that

$$\widehat{\sigma}^{or2}(d^{opt}) = \underbrace{\mathrm{Var}\left(\frac{\mathrm{g}\{A_0, d^{opt}(\boldsymbol{X}_0)\}}{\pi(A_0, \boldsymbol{X}_0)}\{Y_0 - h(A_0, \boldsymbol{X}_0)\} + h\{d^{opt}(\boldsymbol{X}_0), \boldsymbol{X}_0\}\right)}_{\widetilde{\sigma}_0^2(d^{opt};\pi,h)} + o_p(1).$$

Thus, the length of (15) satisfies

$$\sqrt{n}L\{\widehat{V}^{or}(d^{opt}), \alpha\} = 2z_{\alpha/2}\widetilde{\sigma}_0(d^{opt}; \pi, h) + o_p(1). \tag{16}$$

Theorem 3.2. *Assume* (8), (11), (16), *(A1)-(A7) hold and* $s_n \to \infty$. *Assume* $\min_{a=0,1}\pi(a, \boldsymbol{x}) \geq 1/4$, $\forall \boldsymbol{x} \in \mathbb{X}$ *with* $\tau(\boldsymbol{x}) = 0$. *Then, we have*

$$\inf_{d^{opt}\in\mathcal{D}^{opt}} nL^2\{\widehat{V}^{or}(d^{opt}), \alpha\} - nL^2(\widehat{V}_B, \alpha) \geq c^{**}z_{\alpha/2}^2\sigma_0^2 Pr\{\tau(\boldsymbol{X}_0) = 0\} + o(1),$$

where

$$c^{**} = \inf_{\substack{a=0,1\\ \boldsymbol{x}\in\mathbb{X}:\tau(\boldsymbol{x})=0}}\left(\frac{3}{\pi(a, \boldsymbol{x})} - \frac{1}{\pi(1-a, \boldsymbol{x})}\right) \geq 0.$$

In randomized studies, we usually have $\pi(1, \boldsymbol{x}) = 1 - \pi(0, \boldsymbol{x}) = \pi^*, \forall \boldsymbol{x} \in \mathbb{X}$ for some constant $\pi^* > 0$. The condition $\min_{a=0,1}\pi(a, \boldsymbol{x}) \geq 1/4$ thus holds if $1/4 \leq \pi^* \leq 3/4$. Theorem 3.2 implies that the proposed CI is asymptotically narrower than (15) in the

nonregular cases. Heuristically, this is due to the subagging procedure, which averages over estimated OTRs in the nonregular cases, resulting in a smoothed treatment regime $p_{s_n}(\cdot)$. To give a more formal explanation, let's assume $\tau(\boldsymbol{x}) = 0$ and $\pi(1, \boldsymbol{x}) = \pi^*$, $\forall \boldsymbol{x}$. In addition, we assume we know the true propensity score and conditional mean functions and set $\widehat{\pi}_{\mathcal{I}} = \pi^*$ and $\widehat{h}_{\mathcal{I}} = h$. Then it follows from Lemma 2.1 that $\mathrm{E}\widehat{V}_{\mathcal{I}_0}(d) = V_0$ for any regime d. By (11), we have

$$n\mathrm{Var}\{\widehat{V}^{or}(d)\} = n\mathrm{Var}\{\widehat{V}_{\mathcal{I}_0}(d)\} = \sigma_0^2 \mathrm{E}\left(\frac{d^2(\boldsymbol{X}_0)}{\pi^*} + \frac{\{1-d(\boldsymbol{X}_0)\}^2}{1-\pi^*}\right) + \mathrm{Var}\{h(0, \boldsymbol{X}_0)\}$$
$$= \sigma_0^2 \mathrm{E}\left(\frac{d(\boldsymbol{X}_0)}{\pi^*} + \frac{1-d(\boldsymbol{X}_0)}{1-\pi^*}\right) + \mathrm{Var}\{h(0, \boldsymbol{X}_0)\} \geq \sigma_0^2 \min\left(\frac{1}{\pi^*}, \frac{1}{1-\pi^*}\right) + \mathrm{Var}\{h(0, \boldsymbol{X}_0)\}.$$

As for our proposed value estimator, it follows from Condition (A7) that

$$n\mathrm{Var}\{\widehat{V}_B\} \approx \sigma_{s_n}^2 = \sigma_0^2 \mathrm{E}\left(\frac{p_{s_n}^2(\boldsymbol{X}_0)}{\pi^*} + \frac{\{1-p_{s_n}(\boldsymbol{X}_0)\}^2}{1-\pi^*}\right) + \mathrm{Var}\{h(0, \boldsymbol{X}_0)\}$$
$$\approx \sigma_0^2 \left(\frac{1}{4\pi^*} + \frac{1}{4(1-\pi^*)}\right) + \mathrm{Var}\{h(0, \boldsymbol{X}_0)\}.$$

When $1/4 \leq \pi^* \leq 3/4$, we have $1/(4\pi^*) + 1/\{4(1-\pi^*)\} \leq \min\{1/\pi^*, 1/(1-\pi^*)\}$. This implies that our proposed estimator is more efficient than the oracle estimator.

In the regular cases where $\Pr\{\tau(\boldsymbol{X}_0) = 0\} = 0$, we can show $\widehat{V}_B = \widehat{V}^{or}(d^{opt}) + o_p(n^{-1/2})$ and $\widehat{\sigma}_B = \widehat{\sigma}^{or}(d^{opt}) + o_p(1)$ for any $d^{opt} \in \mathcal{D}^{opt}$. This means the proposed CI is asymptotically equivalent to the CI of the oracle method in the regular cases.

4 Multiple time point study

4.1 Optimal dynamic treatment regime

In this section, we consider a multistage study where the treatment decisions are made at a finite number of time points $t_1, \ldots, t_K$. The data for a subject can be summarized as

$$(\boldsymbol{X}_0^{(1)}, A_0^{(1)}, \boldsymbol{X}_0^{(2)}, A_0^{(2)}, \ldots, \boldsymbol{X}_0^{(K)}, A_0^{(K)}, Y_0),$$

where Y_0 denotes the outcome of interest, $\boldsymbol{X}_0^{(1)}$ stands for the set of covariates obtained prior to the time point t_1, $A_0^{(1)}$ denotes the treatment received at t_1. For $k = 2, \ldots, K$, $\boldsymbol{X}_0^{(k)}$ denotes some additional covariates collected between time points t_{k-1} and t_k, and $A_0^{(k)}$ denotes the treatment given at t_k. For simplicity, we assume $A_0^{(1)}, \ldots, A_0^{(K)}$ are all binary treatments. For $k = 1, \ldots, K$, let

$$\bar{\boldsymbol{X}}_0^{(k)} = (\boldsymbol{X}_0^{(1)}, \ldots, \boldsymbol{X}_0^{(k)}) \in \bar{\mathbb{X}}^{(k)} \quad \text{and} \quad \bar{\boldsymbol{A}}_0^{(k)} = (A_0^{(1)}, \ldots, A_0^{(k)}) \in \{0,1\}^k$$

denote a patient's covariates and treatment history. For any $a_1, \ldots, a_K \in \{0,1\}$, denoted by $\bar{\boldsymbol{a}}_k = (a_1, \ldots, a_k)$ for $k = 1, \ldots, K$. The set of all potential outcomes is given by

$$\boldsymbol{W} = \left\{ \left(\boldsymbol{X}_0^{(2)*}(a_1), \boldsymbol{X}_0^{(3)*}(\bar{\boldsymbol{a}}_2), \ldots, \boldsymbol{X}_0^{(K)*}(\bar{\boldsymbol{a}}_{K-1}), Y_0^*(\bar{\boldsymbol{a}}_K) \right) : \forall a_1, \ldots, a_K \in \{0,1\} \right\}, \quad (17)$$

where $\boldsymbol{X}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1})$ denotes the potential time-dependent covariates of a patient that would occur between t_{k-1} and t_k assuming he/she receives treatments $(a_1, \ldots, a_{k-1})$ at decision points $(t_1, \ldots, t_{k-1})$ and $Y_0^*(\bar{\boldsymbol{a}}_K)$ denotes the potential outcome that would result assuming he/she receives treatments $(a_1, \ldots, a_K)$.

A dynamic treatment regime $d = \{d_k\}_{k=1}^K$ is a set of decision rules that treats a patient over time. For $k = 1, \ldots, K$, $d_k = d_k(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k)$ corresponds to the kth decision rule that takes as input a patient's realized covariate and treatment history and outputs a treatment option $a_k \in \{0,1\}$. Let $\bar{d}_k = \{d_j\}_{j=1}^k$ for $k = 1, \ldots, K-1$, the potential outcome associated with d is given by

$$\left(\boldsymbol{X}_0^{(2)*}(d_1), \boldsymbol{X}_0^{(3)*}(\bar{d}_2), \ldots, \boldsymbol{X}_0^{(K)*}(\bar{d}_{K-1}), Y_0^*(d) \right),$$

where $\boldsymbol{X}_0^{(k)*}(\bar{d}_{k-1})$ stands for the potential covariates of a patient between t_{k-1} and t_k assuming he/she receives the treatments sequentially according to the decision rules $(d_1, \ldots, d_{k-1})$ and $Y_0^*(d)$ stands for the potential outcome assuming the treatments he/she receives are determined by the treatment regime d. An optimal dynamic treatment regime d^{opt} is defined

to maximize the average potential outcome, i.e,

$$d^{opt} = \arg\max_{d} \mathrm{E}Y_0^*(d).$$

For any $\bar{\boldsymbol{a}}_K \in \{0,1\}^K$ and $\bar{\boldsymbol{x}}_K \in \bar{\mathbb{X}}^{(K)}$, let $h_K(\bar{\boldsymbol{a}}_K, \bar{\boldsymbol{x}}_K) = \mathrm{E}(Y_0 | \bar{\boldsymbol{X}}_0^{(K)} = \bar{\boldsymbol{x}}_K, \bar{\boldsymbol{A}}_0^{(K)} = \bar{\boldsymbol{a}}_K)$ and $\tau_K(\bar{\boldsymbol{a}}_{K-1}, \bar{\boldsymbol{x}}_K) = h_K\{(\bar{\boldsymbol{a}}_{K-1}, 1), \bar{\boldsymbol{x}}_K\} - h_K\{(\bar{\boldsymbol{a}}_{K-1}, 0), \bar{\boldsymbol{x}}_K\}$. In addition, for $k = K-1, \cdots, 2$, we sequentially define

$$h_k(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{x}}_k) = \mathrm{E}\left(\arg\max_{a_{k+1} \in \{0,1\}} h_{k+1}\{(\bar{\boldsymbol{a}}_k, a_{k+1}), \bar{\boldsymbol{X}}_0^{(k+1)}\} \middle| \bar{\boldsymbol{X}}_0^{(k)} = \bar{\boldsymbol{x}}_k, \bar{\boldsymbol{A}}_0^{(k)} = \bar{\boldsymbol{a}}_k \right),$$

and $\tau_k(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k) = h_k\{(\bar{\boldsymbol{a}}_{k-1}, 1), \bar{\boldsymbol{x}}_k\} - h_k\{(\bar{\boldsymbol{a}}_{k-1}, 0), \bar{\boldsymbol{x}}_k\}$, for any $\bar{\boldsymbol{a}}_k \in \{0,1\}^k$ and $\bar{\boldsymbol{x}}_k \in \bar{\mathbb{X}}^{(k)}$. For $k = 1$, let

$$h_1(a_1, \boldsymbol{x}_1) = \mathrm{E}\left(\arg\max_{a_2 \in \{0,1\}} h_2\{(a_1, a_2), \bar{\boldsymbol{X}}_0^{(2)}\} \middle| \boldsymbol{X}_0^{(1)} = \boldsymbol{x}_1, A_0^{(1)} = a_1 \right)$$

and $\tau_1(\boldsymbol{x}_1) = h_1(1, \boldsymbol{x}_1) - h_1(0, \boldsymbol{x}_1)$ for any $a_1 \in \{0,1\}, \boldsymbol{x}_1 \in \bar{\mathbb{X}}_1$. Under the following two conditions,

(C1.) $Y_0 = \sum_{\bar{\boldsymbol{a}}_K \in \{0,1\}^K} Y_0^*(\bar{\boldsymbol{a}}_K) \mathbb{I}(\bar{\boldsymbol{A}}_0^{(K)} = \bar{\boldsymbol{a}}_K)$ and $\boldsymbol{X}_0^{(k)} = \sum_{\bar{\boldsymbol{a}}_{k-1} \in \{0,1\}^{k-1}} \boldsymbol{X}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1}) \mathbb{I}(\bar{\boldsymbol{A}}_0^{(k-1)} = \bar{\boldsymbol{a}}_{k-1})$, $\forall k = 2, \ldots, K$ and $a_1, \ldots, a_K \in \{0,1\}$,

(C2.) $A_0^{(k)} \perp\!\!\!\perp \boldsymbol{W} | \bar{\boldsymbol{X}}_0^{(k)}, \bar{\boldsymbol{A}}_0^{(k-1)}, \forall k = 1, \ldots, K$ where $\boldsymbol{W}$ is defined in (17),

we can show

$$h(\bar{\boldsymbol{a}}_K, \bar{\boldsymbol{x}}_K) = \mathrm{E}\{Y_0^*(\bar{\boldsymbol{a}}_K) | \bar{\boldsymbol{X}}_0^{(K)*}(\bar{\boldsymbol{a}}_{K-1}) = \bar{\boldsymbol{x}}_K\}, \tag{18}$$

$$h(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{x}}_k) = \mathrm{E}[V_0^{(k+1)}\{\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{X}}_0^{(k+1)*}(\bar{\boldsymbol{a}}_k)\} | \bar{\boldsymbol{X}}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1}) = \bar{\boldsymbol{x}}_k], k = 2, \ldots, K-1, \tag{19}$$

$$h(a_1, \boldsymbol{x}_1) = \mathrm{E}[V_0^{(2)}\{a_1, \bar{\boldsymbol{X}}_0^{(2)*}(a_1)\} | \boldsymbol{X}_0^{(1)} = \boldsymbol{x}_1], \tag{20}$$

where

$$V_0^{(K)}(\bar{\boldsymbol{a}}_{K-1}, \bar{\boldsymbol{x}}_K) = \max_{a_K \in \{0,1\}} \mathrm{E}\{Y_0^*(\bar{\boldsymbol{a}}_K) | \bar{\boldsymbol{X}}_0^{(K)*}(\bar{\boldsymbol{a}}_{K-1}) = \bar{\boldsymbol{x}}_K\},$$
$$V_0^{(k)}(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k) = \max_{a_k \in \{0,1\}} \mathrm{E}[V_0^{(k+1)}\{\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{X}}_0^{(k+1)*}(\bar{\boldsymbol{a}}_k)\} | \bar{\boldsymbol{X}}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1}) = \bar{\boldsymbol{x}}_k], \quad k = 2, \ldots, K-1,$$
$$\bar{\boldsymbol{X}}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1}) = \{\boldsymbol{X}_0^{(1)}, \boldsymbol{X}_0^{(2)*}(a_1), \ldots, \boldsymbol{X}_0^{(k)*}(\bar{\boldsymbol{a}}_{k-1})\}, \quad k = 2, \ldots, K-1.$$

Here, Condition (C2) automatically holds in sequentially randomized studies (Murphy, 2005).

Define the set of dynamic treatment regimes $\mathcal{D}^{opt}$ such that any $d = \{d_k\}_{k=1}^K \in \mathcal{D}^{opt}$ shall satisfy

$$d_K(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k) \in \arg\max_{a \in \{0,1\}} a\tau_k(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k), k = 2, \ldots, K, \quad \text{and} \quad d_1(\boldsymbol{x}_1) \in \arg\max_{a \in \{0,1\}} a\tau_1(\boldsymbol{x}_1), \quad (21)$$

for any $\bar{\boldsymbol{x}}_K \in \bar{\mathbb{X}}^{(K)}, \ldots, \bar{\boldsymbol{x}}_2 \in \bar{\mathbb{X}}^{(2)}, \boldsymbol{x}_1 \in \bar{\mathbb{X}}^{(1)}$ and $\bar{\boldsymbol{a}}_{K-1} \in \{0,1\}^{K-1}, \ldots, \bar{\boldsymbol{a}}_2 \in \{0,1\}^2, a_1 \in \{0,1\}$. By (18)-(20) and backward induction, we can show that

$$\mathcal{D}^{opt} \subseteq \arg\max_d \mathrm{E}Y_0^*(d).$$

Notice that the argmax in (21) is not unique when $\tau_k(\bar{\boldsymbol{a}}_{k-1}, \bar{\boldsymbol{x}}_k) = 0$ or $\tau_1(\boldsymbol{x}_1) = 0$. Therefore, the optimal dynamic treatment regime may not be unique.

4.2 Confidence interval for the optimal value function

In this section, we focus on constructing CIs for the optimal value function $V_0 = \max_d \mathrm{E}Y_0^*(d)$, based on the observed dataset:

$$\left\{O_i = \left(\boldsymbol{X}_i^{(1)}, A_i^{(1)}, \boldsymbol{X}_i^{(2)}, A_i^{(2)}, \ldots, \boldsymbol{X}_i^{(K)}, A_i^{(K)}, Y_i\right) : i = 1, \ldots, n\right\}.$$

For $k = 1, \ldots, K, i = 0, 1, \ldots, n$, let

$$\bar{\boldsymbol{X}}_i^{(k)} = (\boldsymbol{X}_i^{(1)}, \ldots, \boldsymbol{X}_i^{(k)}) \quad \text{and} \quad \bar{\boldsymbol{A}}_i^{(k)} = (A_i^{(1)}, \ldots, A_i^{(k)}).$$

Define the propensity score function $\pi_k(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{x}}_k) = \Pr(A_0^{(k)} = a_k | \bar{\boldsymbol{X}}_0^{(k)} = \bar{\boldsymbol{x}}_k, \bar{\boldsymbol{A}}_0^{(k-1)} = \bar{\boldsymbol{a}}_{k-1})$ for $k = 2, \ldots, K$ and $\pi_1(a_1, \boldsymbol{x}_1) = \Pr(A_0^{(1)} = a_k | \boldsymbol{X}_0^{(1)} = \boldsymbol{x}_1)$. For any dynamic treatment regime $d = \{d_k\}_{k=1}^K$, define

$$\begin{aligned}\widehat{V}_i^{(K)}(d; \pi^*, h^*) &= \frac{\mathrm{g}\{A_i^{(K)}, d_K(\bar{\boldsymbol{A}}_i^{(K-1)}, \bar{\boldsymbol{X}}_i^{(K)})\}}{\pi_K^*(\bar{\boldsymbol{A}}_i^{(K)}, \bar{\boldsymbol{X}}_i^{(K)})}\{Y_i - h_K^*(\bar{\boldsymbol{A}}_i^{(K)}, \bar{\boldsymbol{X}}_i^{(K)})\} \\ &+ h_K^*[\{\bar{\boldsymbol{A}}_i^{(K-1)}, d_K(\bar{\boldsymbol{A}}_i^{(K-1)}, \bar{\boldsymbol{X}}_i^{(K)})\}, \bar{\boldsymbol{X}}_i^{(K)}],\end{aligned}$$

and

$$\begin{aligned}\widehat{V}_i^{(k)}(d; \pi^*, h^*) &= \frac{\mathrm{g}\{A_i^{(k)}, d_k(\bar{\boldsymbol{A}}_i^{(k-1)}, \bar{\boldsymbol{X}}_i^{(k)})\}}{\pi_k^*(\bar{\boldsymbol{A}}_i^{(k)}, \bar{\boldsymbol{X}}_i^{(k)})}\{\widehat{V}_i^{(k+1)}(d; \pi^*, h^*) - h_k^*(\bar{\boldsymbol{A}}_i^{(k)}, \bar{\boldsymbol{X}}_i^{(k)})\} \\ &+ h_k^*[\{\bar{\boldsymbol{A}}_i^{(k-1)}, d_k(\bar{\boldsymbol{A}}_i^{(k-1)}, \bar{\boldsymbol{X}}_i^{(k)})\}, \bar{\boldsymbol{X}}_i^{(k)}],\end{aligned}$$

for $k = K-1, \ldots, 2$, $i = 0, 1, \ldots, n$, where $\pi^* \equiv \{\pi_k^*\}_{k=1}^K$ and $h^* \equiv \{h_k^*\}_{k=1}^K$ denote the estimated propensity score and conditional mean functions. For any $\mathcal{I} \subseteq \mathcal{I}_0$, define the following augmented propensity-score weighted estimator,

$$\widehat{V}_{\mathcal{I}}(d; \pi^*, h^*) = \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|}\left(\frac{\mathrm{g}\{A_i^{(1)}, d_1(\bar{\boldsymbol{X}}_i^{(1)})\}}{\pi_1^*(A_i^{(1)}, \bar{\boldsymbol{X}}_i^{(1)})}\{\widehat{V}_i^{(2)}(d; \pi^*, h^*) - h_1^*(A_i^{(1)}, \bar{\boldsymbol{X}}_i^{(1)})\} + h_1^*\{d_1(\bar{\boldsymbol{X}}_i^{(1)}), \bar{\boldsymbol{X}}_i^{(1)}\}\right).$$

Notice that $\mathrm{E}\widehat{V}_{\mathcal{I}}(d^{opt}; \pi^*, h^*) = V_0$ when either $\pi^* = \pi$ or $h^* = h$.

For any $\mathcal{I} \subseteq \mathcal{I}_0$, let $\{\widehat{\tau}_{\mathcal{I},k}\}_{k=1}^K$, $\{\widehat{h}_{\mathcal{I},k}\}_{k=1}^K$, $\{\widehat{\pi}_{\mathcal{I},k}\}_{k=1}^K$ denote some consistent estimators for $\{\tau_k\}_{k=1}^K$, $\{h_k\}_{k=1}^K$ and $\{\pi_k\}_{k=1}^K$, computed based on the sub-dataset $\{O_i\}_{i \in \mathcal{I}}$. Define the estimated treatment regime $\widehat{d}_{\mathcal{I},k}(\cdot, \cdot) = \mathbb{I}\{\widehat{\tau}_{\mathcal{I},k}(\cdot, \cdot) > 0\}$, $k = 2, \ldots, K$ and $\widehat{d}_{\mathcal{I},1}(\cdot) = \mathbb{I}\{\widehat{\tau}_{\mathcal{I},1}(\cdot) > 0\}$. Define the set

$$\mathcal{S}_{N_0, s_n} = \left\{\mathcal{I} \subseteq \mathcal{I}_0 : |\mathcal{I}| = s_n, \min_{a_1, \ldots, a_K \in \{0,1\}} \sum_{i \in \mathcal{I}} \mathbb{I}(A_i^{(1)} = a_1, \ldots, A_i^{(K)} = a_K) \geq N_0\right\},$$

for some integers $s_n > N_0 > 0$. We summarize our procedure in the following algorithm.

Step 1 Input observations $\{O_i\}_{i \in \mathcal{I}_0}$, $0 < \alpha < 1$ and integers s_n, N_0 and B.

Step 2 For $b = 1, \ldots, B$,

(i) Draw a subset $\mathcal{I}_b$ uniformly from $\mathcal{S}_{N_0,s_n}$.

(ii) Randomly partition $\mathcal{I}_b^c$ into 2 disjoint subsets $\mathcal{I}_b^{c(1)}$ and $\mathcal{I}_b^{c(2)}$ of equal sizes t_n.

(iii) For $j = 1, 2$, let $\mathcal{I}_b^{(j)} = \mathcal{I}_b \cup \mathcal{I}_b^{c(j)}$. Obtain the estimators $\widehat{d}_{\mathcal{I}_b} = \{\widehat{d}_{\mathcal{I}_b,k}\}_{k=1}^K$, $\widehat{\pi}_{\mathcal{I}_b^{(1)}} = \{\widehat{\pi}_{\mathcal{I}_b^{(1)},k}\}_{k=1}^K$, $\widehat{\pi}_{\mathcal{I}_b^{(2)}} = \{\widehat{\pi}_{\mathcal{I}_b^{(2)},k}\}_{k=1}^K$, $\widehat{h}_{\mathcal{I}_b^{(1)}} = \{\widehat{h}_{\mathcal{I}_b^{(1)},k}\}_{k=1}^K$ and $\widehat{h}_{\mathcal{I}_b^{(2)}} = \{\widehat{\pi}_{\mathcal{I}_b^{(2)},k}\}_{k=1}^K$.

Step 3 Compute

$$\widehat{V}_B = \frac{1}{2B}\sum_{b=1}^{B}\left(\widehat{V}_{\mathcal{I}_b^{c(2)}}(\widehat{d}_{\mathcal{I}_b};\widehat{\pi}_{\mathcal{I}_b^{(1)}},\widehat{h}_{\mathcal{I}_b^{(1)}}) + \widehat{V}_{\mathcal{I}_b^{c(1)}}(\widehat{d}_{\mathcal{I}_b};\widehat{\pi}_{\mathcal{I}_b^{(2)}},\widehat{h}_{\mathcal{I}_b^{(2)}})\right),$$

and $\widehat{\sigma}_B^2 = \widehat{s.e}^2(\{\widehat{V}^{(i)}\}_{i=1}^n)$ where

$$\widehat{V}^{(i)} = \frac{1}{n^{(i)}}\sum_{j=1}^{2}\sum_{b=1}^{B}\widehat{V}_{\{i\}}(\widehat{d}_{\mathcal{I}_b};\widehat{\pi}_{\mathcal{I}_b^{(j)}},\widehat{h}_{\mathcal{I}_b^{(j)}})\mathbb{I}(i \notin \mathcal{I}_b^{(j)}),$$

and $n^{(i)} = \sum_{b=1}^{B}\mathbb{I}(i \notin \mathcal{I}_b)$.

Step 4 Output

$$\left[\widehat{V}_B - \frac{z_{\alpha/2}\widehat{\sigma}_B}{\sqrt{n}}, \widehat{V}_B + \frac{z_{\alpha/2}\widehat{\sigma}_B}{\sqrt{n}}\right]. \tag{22}$$

To show the validity of the CI in (22), we introduce the following conditions.

(C3) Assume there exists some positive constant c_0 such that

$$\min_{k=2,\ldots,K}\inf_{\substack{\bar{\boldsymbol{x}}_k\in\bar{\mathbb{X}}^{(k)}\\ \bar{\boldsymbol{a}}_k\in\{0,1\}^k}}\pi_k(\bar{\boldsymbol{a}}_k,\bar{\boldsymbol{x}}_k) \geq c_0 \quad \text{and} \quad \inf_{\substack{\boldsymbol{x}_1\in\bar{\mathbb{X}}^{(1)}\\ a_1=0,1}}\pi_1(a_1,\boldsymbol{x}_1) \geq c_0.$$

(C4) Assume $\sup_{\bar{\boldsymbol{x}}_K\in\bar{\mathbb{X}}^{(K)},a_1,\ldots,a_K\in\{0,1\}}\mathrm{E}[\{Y_0^*(\bar{\boldsymbol{a}}_K)\}^2|\bar{\boldsymbol{X}}_0^{(K)*}(\bar{\boldsymbol{a}}_{K-1}) = \bar{\boldsymbol{x}}_K] = O(1)$.

(C5) Assume there exist some positive constants $\bar{c}$, α and δ_0 such that

$$\max_{k=2,\ldots,K}\Pr(0 < |\tau_k(\bar{\boldsymbol{A}}_0^{(k-1)},\bar{\boldsymbol{X}}_0^{(k)})| < t) \leq \bar{c}t^\alpha \text{ and } \Pr(0 < |\tau_1(\boldsymbol{X}_0^{(1)})| < t) \leq \bar{c}t^\alpha, \quad \forall 0 < t \leq \delta_0.$$

(C6) Assume there exists some constant $\kappa_0 > (\alpha+2)/(2\alpha+2)$ such that

$$\max_{k=2,\ldots,K} \mathrm{E}|\widehat{\tau}_{\mathcal{I},k}(\bar{\boldsymbol{A}}_0^{(k-1)}, \bar{\boldsymbol{X}}_0^{(k)}) - \tau_k(\bar{\boldsymbol{A}}_0^{(k-1)}, \bar{\boldsymbol{X}}_0^{(k)})|^2 = O(|\mathcal{I}|^{-\kappa_0}),$$
$$\mathrm{E}|\widehat{\tau}_{\mathcal{I},1}(\boldsymbol{X}_0^{(1)}) - \tau_1(\boldsymbol{X}_0^{(1)})|^2 = O(|\mathcal{I}|^{-\kappa_0}),$$

for any $\mathcal{I} \subseteq \mathcal{I}_0$.

Notice that Conditions (C3)-(C6) are very similar to (A3)-(A6) in single-stage studies. For any $\mathcal{I} \subseteq \mathcal{I}_0$ with $|\mathcal{I}| = s_n$, define

$$\sigma_{s_n}^2 = \mathrm{Var}\left\{ \mathrm{E}\left(\frac{\mathrm{g}\{A_0^{(1)}, \widehat{d}_{\mathcal{I},1}(\bar{\boldsymbol{X}}_0^{(1)})\}}{\pi_1(A_0^{(1)}, \bar{\boldsymbol{X}}_0^{(1)})} \{\widehat{V}_0^{(2)}(\widehat{d}_{\mathcal{I}}; \pi, h) - h_1(A_0^{(1)}, \bar{\boldsymbol{X}}_0^{(1)})\} + h_1\{\widehat{d}_{\mathcal{I},1}(\bar{\boldsymbol{X}}_0^{(1)}), \bar{\boldsymbol{X}}_0^{(1)}\} \middle| O_0 \right) \right\},$$

where $O_0 = (\bar{\boldsymbol{A}}_0^{(K)}, \bar{\boldsymbol{X}}_0^{(K)}, Y_0)$. We have the following results.

Theorem 4.1. *Assume (C1)-(C6) hold, and s_n satisfies $s_n \asymp n^{\beta_0}$ for some $(2+\alpha)/\{\kappa_0(2+2\alpha)\} < \beta_0 < 1$. Assume $B \gg n$, $\liminf_n \sigma_{s_n} > 0$,*

$$Pr\left(\left\{ \min_{\substack{k=2,\ldots,K \\ \mathcal{I}\subseteq\mathcal{I}_0}} \inf_{\substack{\bar{\boldsymbol{x}}_k \in \bar{\mathbb{X}}^{(k)} \\ \bar{\boldsymbol{a}}_k \in \{0,1\}^k}} \widehat{\pi}_{\mathcal{I},k}(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{x}}_k) \geq c^* \right\} \bigcap \left\{ \inf_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ \boldsymbol{x}_1 \in \bar{\mathbb{X}}^{(1)}, a_1 = 0,1}} \widehat{\pi}_{\mathcal{I},1}(a_1, \boldsymbol{x}_1) \geq c^* \right\} \right) = 1, \quad (23)$$

for some constant $c^* > 0$. In addition, for any $\mathcal{I} \subseteq \mathcal{I}_0$, assume

$$\max_{a=0,1} \max_{k=1,\ldots,K} E|\widehat{\pi}_{\mathcal{I},k}\{(\bar{\boldsymbol{A}}_0^{(k-1)}, a), \bar{\boldsymbol{X}}_0^{(k)}\} - \pi_k\{(\bar{\boldsymbol{A}}_0^{(k-1)}, a), \bar{\boldsymbol{X}}_0^{(k)}\}|^2 = o(|\mathcal{I}|^{-1/2}), \quad (24)$$
$$\max_{a=0,1} \max_{k=1,\ldots,K} E|\widehat{h}_{\mathcal{I},k}\{(\bar{\boldsymbol{A}}_0^{(k-1)}, a), \bar{\boldsymbol{X}}_0^{(k)}\} - h_k\{(\bar{\boldsymbol{A}}_0^{(k-1)}, a), \bar{\boldsymbol{X}}_0^{(k)}\}|^2 = o(|\mathcal{I}|^{-1/2}), \quad (25)$$

where $\bar{\boldsymbol{A}}_0^{(0)} = \emptyset$. Then, we have

$$\frac{\sqrt{n}(\widehat{V}_B - V_0)}{\widehat{\sigma}_B} \xrightarrow{d} N(0,1).$$

According to Theorem 4.1, we can show $\widehat{V}_B - z_\alpha \widehat{\sigma}_B/\sqrt{n}$ is an asymptotic $1-\alpha$ lower confidence bound for V_0.

5 Simulations

5.1 Point treatment study

We consider simulation studies based on the following model:

$$Y_0 = \Phi(X_{0,1}, X_{0,2}) + A_0\tau(X_{0,1}, X_{0,2}) + e_0,$$

where the covariate $X_{0,1}$ and the treatment A_0 are generated from Ber(0.5) and Ber($0.5 + X_{0,1}$), respectively, where Ber(p_0) stands for the Bernoulli distribution with probability p_0. The random error term e_0 satisfies $\mathrm{E}(e_0|A_0, X_{0,1}, X_{0,2}) = 0$. We consider six scenarios. In Scenario (A) and (B), $X_{0,2}$ is generated from Ber(0.5), and

$$e_0 \sim \mathrm{Ber}\{\Phi(X_{0,1}, X_{0,2}) + A_0\tau(X_{0,1}, X_{0,2})\} - \Phi(X_{0,1}, X_{0,2}) - A_0\tau(X_{0,1}, X_{0,2}).$$

In Scenario (C)-(F), $X_{0,2}$ follows a uniform distribution on the interval $[-2, 2]$, and $e_0 \sim N(0, 0.25)$ is independent of A_0, $X_{0,1}$ and $X_{0,2}$. In addition, $X_{0,1}$ and $X_{0,2}$ are independently generated in all scenarios. Table 1 summarizes the information of the baseline function, the contrast function and the optimal value V_0 under different scenarios. In all scenarios, V_0 can be explicitly calculated. The OTR is not uniquely defined in Scenario (A), (C) and (E), since the contrast functions in these scenarios satisfy

$$\Pr\{\tau(X_{0,1}, X_{0,2}) = 0\} = \Pr(X_{0,1} = 0) = \frac{1}{2}.$$

On the contrary, we have $\Pr\{\tau(X_{0,1}, X_{0,2}) = 0\} = 0$ in the remaining three scenarios. For each scenario, we further consider two different sample sizes, $n = 500$ and $n = 1000$. This yields a total of 12 settings.

Comparison is made among the following three methods:

(i) The proposed CI in (5).

(ii) The CI constructed by the online one-step method (12).

(iii) The CI constructed by the oracle method (15) with $d^{opt} = d^{opt,0}$ (see (2)). (Notice that

Table 1: Simulation setting

	(A)	(B)	(C)	(D)	(E)	(F)
$\Phi(x_1, x_2)$	0.3	0.3	x_2^2	x_2^2	x_2^2	x_2^2
$\tau(x_1, x_2)$	$0.4\mathbb{I}(x_1 = 0)$	0.4	$x_1 x_2^2$	$x_2^2 - 4/3$	$2x_1 \cos(\pi x_2/4)$	$2\cos(\pi x_2/4) - 4/\pi$
V_0	0.5	0.7	2	1.85	1.97	1.60

$d^{opt,0}$ is unknown in practice, we implement this method for comparison purposes only.)

All three methods require estimation of the propensity score and conditional mean functions. For scenario (A) and (B), we use the nonparametric maximum likelihood estimator to estimate these functions. For scenario (C)-(F), we estimate these functions using cubic B-splines. More specifically, for $a = 0, 1$, define

$$\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{\pi,a} = \arg\min_{\boldsymbol{\xi}} \sum_{i\in\mathcal{I}} \left(A_i - \sum_{j=1}^{K+4} N_j(X_{i,2})\xi_j \right)^2 \mathbb{I}(X_{i,1} = a),$$

and

$$\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{h_1,a} = \arg\min_{\boldsymbol{\xi}} \sum_{i\in\mathcal{I}} \left(Y_i - \sum_{j=1}^{K+4} N_j(X_{i,2})\xi_j \right)^2 \mathbb{I}(A_i = 1, X_{i,1} = a),$$
$$\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{h_0,a} = \arg\min_{\boldsymbol{\xi}} \sum_{i\in\mathcal{I}} \left(Y_i - \sum_{j=1}^{K+4} N_j(X_{i,2})\xi_j \right)^2 \mathbb{I}(A_i = 0, X_{i,1} = a),$$

where $N_1(\cdot), \ldots, N_{K+4}(\cdot)$ stand for the cubic B-spline basis, and K denotes the number of interior knots. Given K, the interior knots are placed at equally spaced sample quantiles of $\{X_{i,2}\}_{i\in\mathcal{I}_0}$. The hyperparameter K is selected via 5-fold cross-validation. After computing $\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{\pi,a}$, $\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{h_1,a}$ and $\widehat{\boldsymbol{\xi}}_{\mathcal{I}}^{h_0,a}$, we set

$$\widehat{\pi}_{\mathcal{I}}(1, \boldsymbol{x}_1) = \min\left(\sum_{\substack{a=\{0,1\} \\ 1\le j\le K+4}} \mathbb{I}(x_{1,1} = a) N_j(x_{1,2}) \widehat{\xi}_{\mathcal{I},j}^{\pi,a}, 0.05 \right), \widehat{\pi}_{\mathcal{I}}(0, \boldsymbol{x}_1) = \min\{1 - \widehat{\pi}_{\mathcal{I}}(1, \boldsymbol{x}_1), 0.05\},$$
$$\widehat{h}_{\mathcal{I}}(1, \boldsymbol{x}_1) = \sum_{a=0,1} \mathbb{I}(x_{1,1} = a) \sum_{j=1}^{K+4} N_j(x_{1,2}) \widehat{\xi}_{\mathcal{I},j}^{h_1,a}, \widehat{h}_{\mathcal{I}}(0, \boldsymbol{x}_1) = \sum_{a=0,1} \mathbb{I}(x_{1,1} = a) \sum_{j=1}^{K+4} N_j(x_{1,2}) \widehat{\xi}_{\mathcal{I},j}^{h_0,a},$$

Table 2: ACP and AL of the CIs with standard errors in parenthesis

Setting (A)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	93.6 (0.8)	11.1 (0.02)	94.1 (0.7)	12.8 (0.02)	94.0 (0.8)	13.1 (0.02)
1000	93.7 (0.8)	7.8 (0.01)	93.9 (0.8)	8.8 (0.01)	94.1 (0.7)	9.0 (0.01)
Setting (B)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	95.1 (0.7)	11.1 (0.01)	93.9 (0.8)	11.5 (0.02)	95.4 (0.7)	11.2 (0.01)
1000	95.3 (0.7)	7.8 (0.01)	95.5 (0.7)	7.9 (0.01)	95.3 (0.7)	7.8 (0.01)
Setting (C)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	94.1 (0.7)	36.8 (0.04)	92.7 (0.8)	41.1 (0.06)	94.3 (0.7)	38.0 (0.08)
1000	93.9 (0.7)	25.9 (0.02)	93.4 (0.8)	27.4 (0.03)	94.5 (0.7)	26.3 (0.02)
Setting (D)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	95.1 (0.7)	36.6 (0.04)	93.2 (0.8)	40.6 (0.05)	93.7 (0.8)	38.0 (0.20)
1000	94.9 (0.7)	25.7 (0.02)	93.1 (0.8)	27.1 (0.02)	94.2 (0.7)	25.9 (0.02)
Setting (E)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	94.7 (0.7)	22.6 (0.02)	88.2 (1.0)	25.8 (0.03)	94.0 (0.8)	24.6 (0.09)
1000	95.5 (0.7)	15.9 (0.01)	91.9 (0.9)	17.2 (0.01)	95.3 (0.7)	16.7 (0.02)
Setting (F)	Proposed		Online		Oracle	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
500	92.4 (0.8)	21.3 (0.04)	87.5 (1.0)	23.3 (0.03)	93.8 (0.8)	24.3 (0.36)
1000	94.3 (0.7)	14.8 (0.01)	90.8 (0.9)	15.6 (0.01)	94.1 (0.7)	15.3 (0.03)

where $\boldsymbol{x}_1 = (x_{1,1}, x_{1,2})$. The estimated contrast function is defined as

$$\widehat{\tau}_{\mathcal{I}}(\boldsymbol{x}_1) = \widehat{h}_{\mathcal{I}}(1, \boldsymbol{x}_1) - \widehat{h}_{\mathcal{I}}(0, \boldsymbol{x}_1).$$

To calculate the CI in (5), we set $s_n = \lfloor n^{6/7} \rfloor$ where $\lfloor z \rfloor$ denotes the largest integer smaller than or equal to z. Such a choice of s_n satisfies the condition $s_n = o(n)$. To implement the online one-step method, we need to specify l_n. In general, the length of the CI in (12) increases as l_n increases. Nonetheless, l_n should be large enough to guarantee that the bias $V(\widehat{d}_{\mathcal{I}_{l_n}}) - V_0$ is negligible. In Scenario (A) and (B), we set $l_n = 50$. In Scenario (C)-(F), we find that when $l_n = 50$, the resulting CIs have very poor coverage probabilities. Therefore, we set $l_n = 100$ in these scenarios. The variance estimator $\widetilde{\sigma}^2_{\mathcal{I}_{(j)}}$ is computed by

$$\widetilde{\sigma}^2_{\mathcal{I}_{(j)}} = \frac{1}{j(j-1)} \sum_{i=1}^{j} \left(\widehat{V}_{\{i\}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}}) - \widehat{V}_{\mathcal{I}_{(j)}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}}) \right)^2.$$

We implement the simulation program in R. Some subroutines are written in C with the GNU Scientific Library (Galassi et al., 2015) to facilitate the computation.

Reported in Table 2 are the average coverage probability (ACP) and average length (AL) of the CIs in (i)-(iii). Results are aggregated over 1000 replications. It can be seen that all three CIs achieve nominal coverage in Scenario (A)-(D). However, the CIs based on the online one-step method and the oracle method are wider than the proposed CIs in all cases. Take Scenario (A) as an example. ALs of our proposed method are at least 13% smaller than other competing methods. In Scenario (E) and (F), ACPs of the online one-step method are smaller than 90% when $n = 500$. In contrast, ACPs of the proposed CIs are close to the nominal level in all cases. In addition, the proposed CIs achieve smaller ALs in these scenarios.

Notice that in Scenario (B), (D) and (F), the contrast function is almost surely nonzero. In theory, when $l_n = o(n)$, the lengths of all three CIs should be asymptotically the same. However, it can been seen from Table 2 that in finite samples, ALs of the CIs based on our proposed method are always smaller than other competing methods.

Table 3: Simulation setting

	(G)	(H)	(I)
$\Phi(x_{1,1}, x_{1,2}, a_1, x_2)$	$x_{1,1}^2 - a_1(0.25 + x_{1,1}^2)$	x_2^2	0
$\tau(x_{1,1}, x_{1,2}, a_1, x_2)$	$a_1 x_2^2$	0	x_2^2
V_0	1.33	1.58	1.58

5.2 Multiple time point study

Consider the following model:

$$Y_0 = \Phi(X_{0,1}^{(1)}, X_{0,2}^{(1)}, A_0^{(1)}, X_0^{(2)}) + A_0^{(2)}\tau(X_{0,1}^{(1)}, X_{0,2}^{(1)}, A_0^{(1)}, X_0^{(2)}) + e_0^{(2)}, \qquad X_0^{(2)} = A_0^{(1)} X_{0,1}^{(1)} + e_0^{(1)},$$

where $X_{0,1}^{(1)}$ and $X_{0,2}^{(1)}$ are the baseline covariates, $A_0^{(1)}$ and $A_0^{(2)}$ denote the first and second treatment a patient receives at t_1 and t_2, $X_0^{(2)}$ stands for the intermediate covariate collected between t_1 and t_2. Variables $A_0^{(1)}$, $A_0^{(2)}$, $X_{0,1}^{(1)}$, $X_{0,2}^{(1)}$, $e_0^{(1)}$ and $e_0^{(2)}$ are all independent. In addition, we assume $A_0^{(1)}, A_0^{(2)} \sim \text{Ber}(0.5)$, $e_0^{(1)}, e_0^{(2)} \sim N(0, 0.25)$ and $X_{0,1}^{(1)}, X_{0,2}^{(1)} \sim \text{Unif}[-2, 2]$ where $\text{Unif}[a, b]$ denotes the uniform distribution on the interval $[a, b]$.

We consider three scenarios. The functional forms of Φ and τ and the optimal value function V_0 under these scenarios are reported in Table 3. In Scenario (G), we have

$$\begin{aligned} h_1(a_1, \boldsymbol{x}_1) &= \text{E}\{A_0^{(1)} X_0^{(2)} + \Phi(X_{0,1}^{(1)}, X_{0,2}^{(1)}, A_0^{(1)}, X_0^{(2)}) | X_{0,1}^{(1)} = x_{1,1}, X_{0,2}^{(1)} = x_{1,2}, A_0^{(1)} = a_1\} \\ &= a_1(0.25 + x_{1,1}^2) + x_{1,1}^2 - a_1(0.25 + x_{1,1}^2) = x_{1,1}^2, \end{aligned}$$

where $\boldsymbol{x}_1 = (x_{1,1}, x_{1,2})$. Therefore, the first stage contrast function $\tau_1(\cdot)$ equals zero. In Scenario (H), the second stage contrast function $\tau_2(\cdot, \cdot)$ equals zero. Hence, the ODTR is not unique in these two scenarios. In the last scenario, we have

$$h_2(\bar{\boldsymbol{a}}_2, \bar{\boldsymbol{x}}_2) = x_2^2 \quad \text{and} \quad h_1(a_1, \boldsymbol{x}_1) = \text{E}\{(X_{0,2}^{(2)})^2 | \bar{\boldsymbol{X}}_0^{(1)} = \boldsymbol{x}_1, A_0^{(1)} = a_1\} = a_1 x_{1,1}^2 + 0.25,$$

where $\bar{\boldsymbol{x}}_2 = (x_{1,1}, x_{1,2}, x_2)$. In this scenario, the ODTR is uniquely defined and we have $d_1^{opt}(\boldsymbol{x}_1) = 1$, $d_2^{opt}(a_1, \bar{\boldsymbol{x}}_2) = 1$.

We compare our proposed CI (see (22)) with the CI based on the online one-step method,

Table 4: ACP and AL of the CIs with standard errors in parenthesis

Setting (G)	Proposed		Online ($l_n = 200$)		Online ($l_n = 400$)	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
600	93.0 (0.8)	27.4 (0.13)	82.1 (1.2)	38.3 (0.12)	90.8 (0.9)	54.0 (0.18)
1200	92.9 (0.8)	18.2 (0.05)	87.9 (1.0)	24.2 (0.06)	90.7 (0.9)	27.0 (0.07)
Setting (H)	Proposed		Online ($l_n = 200$)		Online ($l_n = 400$)	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
600	92.2 (0.8)	37.2 (0.09)	84.1 (1.2)	45.3 (0.10)	91.4 (0.9)	64.0 (0.13)
1200	92.8 (0.8)	25.4 (0.04)	89.5 (1.0)	28.6 (0.05)	92.8 (0.8)	32.0 (0.05)
Setting (I)	Proposed		Online ($l_n = 200$)		Online ($l_n = 400$)	
n	ACP(%)	AL*100	ACP(%)	AL*100	ACP(%)	AL*100
600	91.8 (0.9)	39.4 (0.14)	84.5 (1.2)	44.5 (0.10)	90.1 (0.9)	63.1 (0.13)
1200	93.0 (0.8)	26.4 (0.04)	90.5 (0.9)	28.4 (0.05)	92.5 (0.8)	31.8 (0.05)

defined as $[\widehat{V}^{on} - z_{\alpha/2}\widehat{\sigma}^{on}/\sqrt{n-l_n}, \widehat{V}^{on} + z_{\alpha/2}\widehat{\sigma}^{on}/\sqrt{n-l_n}]$ where

$$\widehat{V}^{on} = \left(\sum_{j=l_n}^{n-1} \widetilde{\sigma}_{\mathcal{I}_{(j)}}^{-1}\right)^{-1} \left(\sum_{j=l_n}^{n-1} \widetilde{\sigma}_{\mathcal{I}_{(j)}}^{-1} \widehat{V}_{\{j+1\}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}})\right), \quad \widehat{\sigma}^{on} = \left(\sum_{j=l_n}^{n-1} \widetilde{\sigma}_{\mathcal{I}_{(j)}}^{-1}/(n-l_n)\right)^{-1}.$$

$$\widetilde{\sigma}^2_{\mathcal{I}_{(j)}} = \frac{1}{j(j-1)} \sum_{i=1}^{j} \left(\widehat{V}_{\{i\}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}}) - \widehat{V}_{\mathcal{I}_{(j)}}(\widehat{d}_{\mathcal{I}_{(j)}}; \widehat{\pi}_{\mathcal{I}_{(j)}}, \widehat{h}_{\mathcal{I}_{(j)}})\right)^2,$$

for some divergent sequence l_n. Notice that both methods require to calculate $\widehat{h}_{\mathcal{I}} = \{\widehat{h}_{\mathcal{I},k}\}_{k=1}^2$, $\widehat{\pi}_{\mathcal{I}} = \{\widehat{\pi}_{\mathcal{I},k}\}_{k=1}^2$, $\widehat{d}_{\mathcal{I}} = \{\widehat{d}_{\mathcal{I},k}\}_{k=1}^2$. These estimators are computed based on cubic B-spline methods. To save space, we present the detailed estimating procedure in Section S1 of the supplementary article.

We consider two sample sizes, $n = 600$ and $n = 1200$. Similar to Section 5.1, we set $s_n = \lfloor n^{6/7} \rfloor$ when implementing (22). In Table 4, we report the ACP and AL of the proposed CI and the CI based on online one-step method, with $l_n = 200$ and $l_n = 400$. It can be seen that ACPs of our proposed CIs are close to the nominal level in almost all cases. In contrast, ACPs of the CIs based on the online one-step method are well below the nominal level in Scenario (G) and (H). Moreover, CIs based on our proposed method are much shorter than those based on the online one-step method.

Table 5: Estimated value functions and confidence intervals

Method	Estimated value function	95% CI	Length of CI
Proposed	399.5	$[387.9, 411.2]$	23.2
Online ($l_n = 50$)	399.2	$[385.6, 412.7]$	27.1
Online ($l_n = 100$)	398.3	$[384.4, 412.2]$	27.7
Online ($l_n = 200$)	403.9	$[389.5, 418.4]$	28.9

6 Real data analysis

In this section, we apply the proposed method to a data from AIDS Clinical Trials Group Protocol 175 (ACTG175). We focus on a subset of the data which consists of 1046 patients that were treated with either ZDV + zalcitabine (zal) ($A = 0$) or ZDV + didanosine (ddI) ($A = 1$). The outcome of interests were CD4 count (cells/mm^3) at 20 ± 5 weeks after receiving the treatment. Lu et al. (2013) and Fan et al. (2017) found that the age variable has significant interaction with the treatment. Therefore, in the following, we use age to construct the OTR. Since ACTG175 is a randomized trial, the no unmeasured confounders assumption (A2) automatically holds.

In Table 5, we report the estimated optimal value function and its 95% CI based on our proposed method and the online one-step method with $l_n = 50, 100$ and 200. To construct these CIs, we set $\widehat{\pi}_{\mathcal{I}} = 0.5$ for any $\mathcal{I} \subseteq \mathcal{I}_0$. The conditional mean functions are estimated using cubic B-splines. The detailed estimating procedure is very similar to that in Section 5.1 and is hence omitted for brevity. In addition, we set $s_n = \lfloor n^{6/7} \rfloor$, as in simulations.

It can be seen from Table 5 that all methods yield similar estimated optimal value functions. These estimated values are larger than those based on linear decision rules (see Section 4 in Fan et al., 2017). Besides, we notice that our proposed CI is at least 16% shorter compared to those based on the online one-step method. Such phenomenon is consistent with our theoretical findings and simulation results.

A Proof of Theorem 2.1

For any $\mathcal{I} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{I}_0$, the estimated treatment regime $|\widehat{d}_{\mathcal{I}}(\cdot)|$ is upper bounded by 1. It follows from the ANOVA decomposition of Efron and Stein (1981) that

$$\begin{aligned} \widehat{d}_{\mathcal{I}}(\boldsymbol{x}) &= p_s(\boldsymbol{x}) + \sum_{i\in\mathcal{I}} d_{s,1}(O_i;\boldsymbol{x}) + \sum_{\substack{i,j\in\mathcal{I}\\ i\neq j}} d_{s,2}(O_i,O_j;\boldsymbol{x}) \\ &+ \sum_{\substack{i,j,k\in\mathcal{I}\\ i\neq j, i\neq k, j\neq k}} d_{s,3}(O_i,O_j,O_k;\boldsymbol{x}) + \cdots + d_{s,s}(O_{i_1},O_{i_2},\ldots,O_{i_s};\boldsymbol{x}), \quad \forall \boldsymbol{x}, \end{aligned} \tag{26}$$

where $p_s(x) = \mathrm{E}\widehat{d}_{\mathcal{I}}(\boldsymbol{x}) = \Pr(\widehat{d}_{\mathcal{I}}(\boldsymbol{x}) = 1)$, is the grand mean; $d_{s,1}(o;\boldsymbol{x}) = \mathrm{E}\{\widehat{d}_{\mathcal{I}}(\boldsymbol{x})|O_{i_1} = o\} - p_s(\boldsymbol{x})$, is the main effect;

$$d_{s,2}(o_1,o_2;\boldsymbol{x}) = \mathrm{E}\{\widehat{d}_{\mathcal{I}}(\boldsymbol{x})|O_{i_1}=o_1, O_{i_2}=o_2\} - \mathrm{E}\{\widehat{d}_{\mathcal{I}}(\boldsymbol{x})|O_{i_1}=o_1\} - \mathrm{E}\{\widehat{d}_{\mathcal{I}}(\boldsymbol{x})|O_{i_2}=o_2\} + p_s(\boldsymbol{x}),$$

is the second-order interaction; etc.

All the 2^s random variables on the right-hand side (RHS) of (26) are uncorrelated. Therefore,

$$\sum_{k=1}^{n} \binom{s}{k} \mathrm{E} d_{s,k}^2(O_{i_1},O_{i_2},\ldots,O_{i_k};\boldsymbol{x}) = \mathrm{Var}\{\widehat{d}_{\mathcal{I}}(\boldsymbol{x})\} \leq \mathrm{E}\widehat{d}_{\mathcal{I}}^2(\boldsymbol{x}) \leq 1. \tag{27}$$

In the following, we show $\eta_2 = o_p(n^{-1/2})$. By definition, this implies $\widehat{V}_\infty = \eta_1 + o_p(n^{-1/2})$. Notice that

$$\eta_2 = \underbrace{\frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \frac{(2A_i-1)R_{\mathcal{I}}(\boldsymbol{X}_i)}{\pi(A_i,\boldsymbol{X}_i)}\{Y_i - h(A_i,\boldsymbol{X}_i)\}}_{\eta_3} + \underbrace{\frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \tau(\boldsymbol{X}_i) R_{\mathcal{I}}(\boldsymbol{X}_i)}_{\eta_4}.$$

Below, we break the proof into two steps. In the first step, we show $\eta_3 = o_p(n^{-1/2})$. In the second step, we prove $\eta_4 = o_p(n^{-1/2})$.

Step 1: For $i = 0, 1, \ldots, n$, let $\omega_{0,i} = (1-A_i)\{Y_i - h(0,\boldsymbol{X}_i)\}/\pi(0,\boldsymbol{X}_i)$ and $\omega_{1,i} =$

$A_i\{Y_i - h(1, \boldsymbol{X}_i)\}/\pi(1, \boldsymbol{X}_i)$. We have

$$\eta_3 = -\frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \omega_{0,i} R_{\mathcal{I}}(\boldsymbol{X}_i) + \frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \omega_{1,i} R_{\mathcal{I}}(\boldsymbol{X}_i).$$

Below, we show

$$\eta_3^{(1)} \equiv \frac{1}{(n-s_n)\binom{n}{s_n}} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \omega_{0,i} R_{\mathcal{I}}(\boldsymbol{X}_i) = o_p(n^{-1/2}). \tag{28}$$

It follows from (26) that

$$\eta_3^{(1)} = \frac{1}{\binom{n}{s_n}(n-s_n)} \sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0 \\ |\mathcal{I}|=s_n}} \sum_{i\in\mathcal{I}^c} \omega_{0,i} \left(\sum_{k=1}^{s_n} \sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}} d_{s_n,k}(O_{j_1}, \ldots, O_{j_k}; \boldsymbol{X}_i) \right).$$

Notice that $(n-s_n)\binom{n}{s_n} = (n-s_n)\binom{n}{n-s_n} = n\binom{n-1}{n-s_n-1} = n\binom{n-1}{s_n}$. With some calculations, we have

$$\begin{aligned}
\eta_3^{(1)} &= \frac{1}{n} \sum_{i=1}^{n} \omega_{0,i} \sum_{k=1}^{s_n} \frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}} \sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}} d_{s_n,k}(O_{j_1}, \ldots, O_{j_k}; \boldsymbol{X}_i) \\
&= \underbrace{\frac{1}{n} \sum_{i=1}^{n} \omega_{0,i} \sum_{k=1}^{l_0-1} \frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}} \sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}} d_{s_n,k}(O_{j_1}, \ldots, O_{j_k}; \boldsymbol{X}_i)}_{\eta_3^{(2)}} \\
&+ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \omega_{0,i} \sum_{k=l_0}^{s_n} \frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}} \sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}} d_{s_n,k}(O_{j_1}, \ldots, O_{j_k}; \boldsymbol{X}_i)}_{\eta_3^{(3)}},
\end{aligned}$$

for some fixed integer $l_0 \geq 2$ that satisfies $l_0 > 1/(1-\beta_0)$.

By (A1) and (A2), we have for any $i = 1, \ldots, n$,

$$\mathrm{E}(\omega_{0,i}|\boldsymbol{X}_i) = \mathrm{E}\left(\frac{1-A_i}{\pi(0, \boldsymbol{X}_i)} \{Y_i^*(0) - h(0, \boldsymbol{X}_i)\} | \boldsymbol{X}_i \right) = \mathrm{E}[\{Y_i^*(0) - h(0, \boldsymbol{X}_i)\}|\boldsymbol{X}_i] = 0.$$

By Condition (A3) and (A4), we have

$$\begin{aligned}\max_{i\in\mathcal{I}_0}\mathrm{E}(\omega_{0,i}^2|\boldsymbol{X}_i) &\le \max_{i\in\mathcal{I}_0}\mathrm{E}\left(\frac{\{Y_i^*(0)-h(0,\boldsymbol{X}_i)\}^2}{\pi^2(0,\boldsymbol{X}_i)}\middle|\boldsymbol{X}_i\right)\le\max_{i\in\mathcal{I}_0}\frac{1}{c_0^2}\mathrm{E}[\{Y_i^*(0)-h(0,\boldsymbol{X}_i)\}^2|\boldsymbol{X}_i]\\ &\le \max_{i\in\mathcal{I}_0}\frac{1}{c_0^2}\mathrm{E}[\{Y_i^*(0)\}^2|\boldsymbol{X}_i]\le\frac{1}{c_0^2}\sup_{\boldsymbol{x}}\mathrm{E}[\{Y_0^*(0)\}^2|\boldsymbol{X}_0=\boldsymbol{x}]\le\bar{c}_*, \qquad (29)\end{aligned}$$

for some constant $\bar{c}_*>0$. Here, the first inequality in (29) is due to that $\mathrm{E}\{Y_i^*(0)|\boldsymbol{X}_i\}=h(0,\boldsymbol{X}_i)$. Therefore, we have

$$\begin{aligned}n\mathrm{E}(\eta_3^{(3)})^2 &\le \mathrm{E}\sum_{i=1}^n\omega_{0,i}^2\left(\sum_{k=l_0}^{s_n}\frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}d_{s_n,k}(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_i)\right)^2\\ &\le \sum_{i=1}^n\mathrm{E}\left\{\mathrm{E}^{O_i}\omega_{0,i}^2\left(\sum_{k=l_0}^{s_n}\frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}d_{s_n,k}(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_i)\right)^2\right\}\\ &= \sum_{i=1}^n\mathrm{E}\left(\omega_{0,i}^2\mathrm{E}^{O_i}\sum_{k=l_0}^{s_n}\frac{\binom{n-1-k}{s_n-k}^2}{\binom{n-1}{s_n}^2}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}d_{s_n,k}^2(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_i)\right)\\ &= n\mathrm{E}\left(\omega_{0,1}^2\mathrm{E}^{O_1}\sum_{k=l_0}^{s_n}\frac{\binom{n-1-k}{s_n-k}^2}{\binom{n-1}{s_n}^2}\binom{n-1}{k}d_{s_n,k}^2(O_2,\ldots,O_{k+1};\boldsymbol{X}_1)\right)\\ &\le n\max_{k\ge l_0}\frac{\binom{n-1-k}{s_n-k}^2\binom{n-1}{k}}{\binom{n-1}{s_n}^2\binom{s_n}{k}}\mathrm{E}\omega_{0,1}^2=\max_{k\ge l_0}\frac{n\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}}\mathrm{E}(\mathrm{E}^{\boldsymbol{X}_1}\omega_{0,1}^2)\le\bar{c}_*n\frac{s_n^{l_0}}{(n-1)^{l_0}},\end{aligned}$$

where E^{O_i} and $\mathrm{E}^{\boldsymbol{X}_i}$ denote the conditional expectation given O_i and $\boldsymbol{X}_i$, respectively, the first inequality is due to Cauchy-Schwarz inequality, the first equality follows by the fact that $d_{s_n,k_1}(O_{j_1^{(1)}},\cdots,O_{j_{k_1}^{(1)}};\boldsymbol{x})$ and $d_{s_n,k_2}(O_{j_1^{(2)}},\cdots,O_{j_{k_2}^{(2)}};\boldsymbol{x})$ are independent for any $\{j_1^{(1)},\ldots,j_{k_1}^{(1)}\}\neq\{j_1^{(2)},\ldots,j_{k_2}^{(2)}\}$, the third inequality is due to (27) and the last inequality is due to (29).

By the definitions of β_0 and l_0, we have $ns_n^{l_0}/(n-1)^{l_0}\asymp n^{1+\beta_0 l_0}/n^{l_0}=o(1)$. This together with Chebyshev's inequality gives

$$\eta_3^{(3)}=o_p(n^{-1/2}). \qquad (30)$$

In addition, it follows from Cauchy-Schwarz inequality that

$$n\mathrm{E}(\eta_3^{(2)})^2 \le \frac{l_0-1}{n}\sum_{k=1}^{l_0-1}\mathrm{E}\left(\frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}}\sum_{i=1}^{n}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}\omega_{0,i}d_{s_n,k}(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_i)\right)^2.$$

For any $1 \le i^{(1)}, i^{(2)} \le n$ and $\{j_1^{(1)},\ldots,j_k^{(1)}\} \subseteq \mathcal{I}_{(-i^{(1)})}$, $\{j_1^{(2)},\ldots,j_k^{(2)}\} \subseteq \mathcal{I}_{(-i^{(2)})}$, we have

$$\mathrm{E}\omega_{0,i^{(1)}}d_{s_n,k}(O_{j_1^{(1)}},\ldots,O_{j_k^{(1)}};\boldsymbol{X}_{i^{(1)}})\omega_{0,i^{(2)}}d_{s_n,k}(O_{j_1^{(2)}},\ldots,O_{j_k^{(2)}};\boldsymbol{X}_{i^{(2)}}) = 0,$$

for any $\{i^{(1)}, j_1^{(1)},\ldots,j_k^{(1)}\} \neq \{i^{(2)}, j_1^{(2)},\ldots,j_k^{(2)}\}$. Let $\sigma_k^2 = \mathrm{E}\omega_{0,1}^2 d_{s_n,k}^2(O_2,\ldots,O_{k+1};\boldsymbol{X}_1)$. It follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
&|\mathrm{E}\omega_{0,i^{(1)}}d_{s_n,k}(O_{j_1^{(1)}},\ldots,O_{j_k^{(1)}};\boldsymbol{X}_{i^{(1)}})\omega_{0,i^{(2)}}d_{s_n,k}(O_{j_1^{(2)}},\ldots,O_{j_k^{(2)}};\boldsymbol{X}_{i^{(2)}})| \\
\le\ & \frac{1}{2}\mathrm{E}|\omega_{0,i^{(1)}}d_{s_n,k}(O_{j_1^{(1)}},\ldots,O_{j_k^{(1)}};\boldsymbol{X}_{i^{(1)}})|^2 + \frac{1}{2}\mathrm{E}|\omega_{0,i^{(2)}}d_{s_n,k}(O_{j_1^{(2)}},\ldots,O_{j_k^{(2)}};\boldsymbol{X}_{i^{(2)}})|^2 = \sigma_k^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
n\mathrm{E}(\eta_3^{(2)})^2 &\le \frac{l_0-1}{n}\sum_{k=1}^{l_0-1}\frac{\binom{n-1-k}{s_n-k}^2}{\binom{n-1}{s_n}^2}\left(\sum_{i=1}^{n}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}\mathrm{E}\omega_{0,i}^2 d_{s_n,k}(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_i)^2\right. \\
&\left. + \sum_{\substack{1\le i\le n\\ 1\le j\le n\\ i\neq j}}\sum_{\substack{\{i_1,\ldots,i_k\}\subseteq\mathcal{I}_{(-i)}\\ \{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-j)}\\ \{i,i_1,\ldots,i_k\}=\{j,j_1,\ldots,j_k\}}}\mathrm{E}\omega_{0,i}d_{s_n,k}(O_{i_1},\ldots,O_{i_k};\boldsymbol{X}_i)\omega_{0,j}d_{s_n,k}(O_{j_1},\ldots,O_{j_k};\boldsymbol{X}_j)\right) \\
&\le \frac{l_0-1}{n}\sum_{k=1}^{l_0-1}\frac{\binom{n-1-k}{s_n-k}^2}{\binom{n-1}{s_n}^2}\left(\sum_{i=1}^{n}\sum_{\{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-i)}}\sigma_k^2 + \sum_{\substack{1\le i\le n\\ 1\le j\le n\\ i\neq j}}\sum_{\substack{\{i_1,\ldots,i_k\}\subseteq\mathcal{I}_{(-i)}\\ \{j_1,\ldots,j_k\}\subseteq\mathcal{I}_{(-j)}\\ \{i,i_1,\ldots,i_k\}=\{j,j_1,\ldots,j_k\}}}\sigma_k^2\right) \\
&= \frac{l_0-1}{n}\sum_{k=1}^{l_0-1}\frac{\binom{n-1-k}{s_n-k}^2}{\binom{n-1}{s_n}^2}\left\{n\binom{n-1}{k}\sigma_k^2 + kn\binom{n-1}{k}\sigma_k^2\right\} \asymp \sum_{k=1}^{l_0-1}\frac{\binom{n-1-k}{s_n-k}}{\binom{n-1}{s_n}}\binom{s_n}{k}\sigma_k^2,
\end{aligned}$$

where the second inequality is due to Cauchy-Schwarz inequality. By (27) and (29), we

have

$$\begin{aligned}\binom{s_n}{k}\sigma_k^2 &\leq \binom{s_n}{k}\mathrm{E}\omega_1^2 d_{s_n,k}^2(O_2,\ldots,O_{k+1};\boldsymbol{X}_1) \leq \binom{s_n}{k}\mathrm{E}\{\mathrm{E}(\omega_{0,1}^2|\boldsymbol{X}_1)d_{s_n,k}^2(O_2,\ldots,O_{k+1};\boldsymbol{X}_1)\}\\ &\leq \bar{c}_*\binom{s_n}{k}\mathrm{E}\{d_{s_n,k}^2(O_2,\ldots,O_{k+1};\boldsymbol{X}_1)\} \leq \bar{c}_*.\end{aligned}$$

Therefore,

$$n\mathrm{E}(\eta_3^{(2)})^2 \asymp \sum_{k=1}^{l_0-1}(k+1)\frac{s_n^k}{(n-1)^k} \leq l_0^2\frac{s_n}{n-1} \to 0.$$

By Cauchy-Schwarz inequality, we obtain $\eta_3^{(2)} = o_p(n^{-1/2})$. This together with (30) gives $\eta_3^{(1)} = o_p(n^{-1/2})$. Similarly, we can show

$$\frac{1}{(n-s_n)\binom{n}{s_n}}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\omega_{1,i}R_{\mathcal{I}}(\boldsymbol{X}_i) = o_p(n^{-1/2}).$$

This together with (28) proves $\eta_3 = o_p(n^{-1/2})$.

Step 2: Notice that

$$\begin{aligned}\eta_4 &= \frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\mathcal{I}\subseteq\mathcal{I}_0,|\mathcal{I}|=s_n}\sum_{i\in\mathcal{I}^c}\tau(\boldsymbol{X}_i)[\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i) - p_{s_n}(\boldsymbol{X}_i)]\\ &= \underbrace{\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\mathcal{I}\subseteq\mathcal{I}_0,|\mathcal{I}|=s_n}\sum_{i\in\mathcal{I}^c}\tau(\boldsymbol{X}_i)[\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i) - \mathbb{I}\{\tau(\boldsymbol{X}_i)>0\}]}_{\eta_4^{(1)}}\\ &- \underbrace{\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\mathcal{I}\subseteq\mathcal{I}_0,|\mathcal{I}|=s_n}\sum_{i\in\mathcal{I}^c}\tau(\boldsymbol{X}_i)[p_{s_n}(\boldsymbol{X}_i) - \mathbb{I}\{\tau(\boldsymbol{X}_i)>0\}]}_{\eta_4^{(2)}}.\end{aligned}$$

To prove $\eta_4 = o_p(n^{-1/2})$, it suffices to show $\eta_4^{(1)}, \eta_4^{(2)} = o_p(n^{-1/2})$.

With some calculations, we have

$$\begin{aligned}\eta_4^{(1)} &= \underbrace{\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\tau(\boldsymbol{X}_i)[\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i)-\mathbb{I}\{\tau(\boldsymbol{X}_i)>0\}]\mathbb{I}\{|\tau(\boldsymbol{X}_i)|\le\tau_n\}}_{\eta_4^{(3)}}\\ &+ \underbrace{\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\tau(\boldsymbol{X}_i)[\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i)-\mathbb{I}\{\tau(\boldsymbol{X}_i)>0\}]\mathbb{I}\{|\tau(\boldsymbol{X}_i)|>\tau_n\}}_{\eta_4^{(4)}},\end{aligned}$$

for some sequence τ_n that will be specified later.

It follows from Condition (A5) that

$$\begin{aligned}\mathrm{E}|\eta_4^{(3)}| &\le \frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\mathcal{I}\subseteq\mathcal{I}_0,|\mathcal{I}|=s_n}\sum_{i\in\mathcal{I}^c}\mathrm{E}|\tau(\boldsymbol{X}_i)|\mathbb{I}\{|\tau(\boldsymbol{X}_i)|\le\tau_n\} \qquad (31)\\ &= \frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\mathcal{I}\subseteq\mathcal{I}_0,|\mathcal{I}|=s_n}\sum_{i\in\mathcal{I}^c}\mathrm{E}|\tau(\boldsymbol{X}_i)|\mathbb{I}\{0<|\tau(\boldsymbol{X}_i)|\le\tau_n\}\le\bar{c}\tau_n^{1+\alpha},\end{aligned}$$

for any τ_n such that $\tau_n\le\delta_0$. Moreover, since $\widehat{d}_{\mathcal{I}}(\boldsymbol{X}_i)\neq\mathbb{I}\{\tau(\boldsymbol{X}_i)>0\}$ only when $|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_i)-\tau(\boldsymbol{X}_i)|>|\tau(\boldsymbol{X}_i)|$, $\mathrm{E}|\eta_4^{(4)}|$ can be upper bounded by

$$\begin{aligned}&\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\mathrm{E}|\tau(\boldsymbol{X}_i)|\mathbb{I}(|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_i)-\tau(\boldsymbol{X}_i)|>|\tau(\boldsymbol{X}_i)|)\mathbb{I}\{|\tau(\boldsymbol{X}_i)|>\tau_n\}\\ &\le\frac{1}{\binom{n}{s_n}(n-s_n)}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\mathrm{E}|\tau(\boldsymbol{X}_i)|\frac{|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_i)-\tau(\boldsymbol{X}_i)|^2}{|\tau(\boldsymbol{X}_i)|^2}\mathbb{I}\{|\tau(\boldsymbol{X}_i)|>\tau_n\}\\ &\le\frac{1}{\tau_n\binom{n}{s_n}(n-s_n)}\sum_{\substack{\mathcal{I}\subseteq\mathcal{I}_0\\|\mathcal{I}|=s_n}}\sum_{i\in\mathcal{I}^c}\mathrm{E}|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_i)-\tau(\boldsymbol{X}_i)|^2\le\frac{s_n^{-\kappa_0}}{\tau_n}\asymp\frac{1}{\tau_n n^{\beta_0\kappa_0}}, \qquad (32)\end{aligned}$$

where the last inequality is due to Condition (A6). Combining (32) together with (31) gives $\mathrm{E}|\eta_4^{(1)}|=O(\tau_n^{1+\alpha}+\tau_n^{-1}n^{-\beta_0\kappa_0})$. Set $\tau_n=n^{-\beta_0\kappa_0/(2+\alpha)}$, we obtain that

$$\mathrm{E}|\eta_4^{(1)}|=O\left(n^{-\beta_0\kappa_0(1+\alpha)/(2+\alpha)}\right)=o(n^{-1/2}),$$

where the last equality is due to the condition $\beta_0 > (2+\alpha)/\{\kappa_0(1+\alpha)\}$. By Markov's inequality, we obtain $\eta_4^{(1)} = o_p(n^{-1/2})$. As for $\eta_4^{(2)}$, we have

$$\begin{aligned}\eta_4^{(2)} &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\tau(\boldsymbol{X}_i)[p_{s_n}(\boldsymbol{X}_i) - \mathbb{I}\{\tau(\boldsymbol{X}_i) > 0\}]\mathbb{I}\{|\tau(\boldsymbol{X}_i)| \le \tau_n\}}_{\eta_4^{(5)}} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\tau(\boldsymbol{X}_i)[p_{s_n}(\boldsymbol{X}_i) - \mathbb{I}\{\tau(\boldsymbol{X}_i) > 0\}]\mathbb{I}\{|\tau(\boldsymbol{X}_i)| > \tau_n\}}_{\eta_4^{(6)}}.\end{aligned}$$

Similar to (31), we can show

$$\mathrm{E}|\eta_4^{(5)}| \le \bar{c}\tau_n^{1+\alpha}, \tag{33}$$

for any τ_n such that $\tau_n \le \delta_0$.

Besides, it follows from Chebysev's inequality that

$$\begin{aligned}&|p_{s_n}(\boldsymbol{X}_0) - \mathbb{I}\{\tau(\boldsymbol{X}_0) > 0\}| = |\Pr^{\boldsymbol{X}_0}\{\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_0) > 0\} - \mathbb{I}\{\tau(\boldsymbol{X}_0) > 0\}| \\ \le\ &\Pr^{\boldsymbol{X}_0}\{|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_0) - \tau(\boldsymbol{X}_0)| \ge |\tau(\boldsymbol{X}_0)|\} \le \mathrm{E}^{\boldsymbol{X}_0}\frac{|\widehat{\tau}_{\mathcal{I}}(\boldsymbol{X}_0) - \tau(\boldsymbol{X}_0)|^2}{|\tau(\boldsymbol{X}_0)|^2},\end{aligned} \tag{34}$$

where $\Pr^{\boldsymbol{X}_0}(\cdot)$ denotes the conditional probability given $\boldsymbol{X}_0$, and $\mathcal{I}$ is an arbitrary subset of $\{1,\ldots,n\}$ with $|\mathcal{I}| = s_n$. Therefore, using similar arguments in bounding $\mathrm{E}|\eta_4^{(4)}|$, we can show $\mathrm{E}|\eta_4^{(6)}| = O(\tau_n^{-1}n^{-\beta_0\kappa_0})$. Combining this together with (33), we've shown

$$\mathrm{E}|\eta_4^{(2)}| = O\left(\tau_n^{1+\alpha} + \frac{1}{\tau_n n^{\beta_0\kappa_0}}\right).$$

Set $\tau_n = n^{-\beta_0\kappa_0/(2+\alpha)}$, by Markov's inequality, we obtain $\eta_4^{(2)} = o_p(n^{-1/2})$. This proves $\eta_4 = o_p(n^{-1/2})$.

To summarize, we've shown $\eta_2 = o_p(n^{-1/2})$. Next, we show $V_0 = \mathrm{E}\eta_1 + o(n^{-1/2})$. Let $\mathrm{E}^{A_0,\boldsymbol{X}_0}$ denote the conditional expectation given A_0 and $\boldsymbol{X}_0$. It follows from the definitions

of V_0 and η_1 that

$$
\begin{aligned}
& V_0 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(\frac{\mathrm{g}\{A_i, p_{s_n}(\boldsymbol{X}_i)\}}{\pi(A_0, \boldsymbol{X}_i)} \{Y_i - h(A_i, \boldsymbol{X}_i)\} + h(p_{s_n}(\boldsymbol{X}_i), \boldsymbol{X}_i) \right) \\
= \; & V_0 - \mathrm{E}\left(\frac{\mathrm{g}\{A_0, p_{s_n}(\boldsymbol{X}_0)\}}{\pi(A_0, \boldsymbol{X}_0)} \{Y_0 - h(A_0, \boldsymbol{X}_0)\} + h(p_{s_n}(\boldsymbol{X}_0), \boldsymbol{X}_0) \right) \\
= \; & V_0 - \mathrm{E}\left(\frac{\mathrm{g}\{A_0, p_{s_n}(\boldsymbol{X}_0)\}}{\pi(A_0, \boldsymbol{X}_0)} \mathrm{E}^{A_0, \boldsymbol{X}_0} \{Y_0 - h(A_0, \boldsymbol{X}_0)\} + h(p_{s_n}(\boldsymbol{X}_0), \boldsymbol{X}_0) \right) \\
= \; & V_0 - \mathrm{E}[p_{s_n}(\boldsymbol{X}_0) h(1, \boldsymbol{X}_0) + \{1 - p_{s_n}(\boldsymbol{X}_0)\} h(0, \boldsymbol{X}_0)] \\
= \; & \mathrm{E}[h(0, \boldsymbol{X}_0) + \tau(\boldsymbol{X}_0) \mathbb{I}\{\tau(\boldsymbol{X}_0) > 0\}] - \mathrm{E}[p_{s_n}(\boldsymbol{X}_0) h(1, \boldsymbol{X}_0) + \{1 - p_{s_n}(\boldsymbol{X}_0)\} h(0, \boldsymbol{X}_0)] \\
= \; & \mathrm{E}\tau(\boldsymbol{X}_0)[\mathbb{I}\{\tau(\boldsymbol{X}_0) > 0\} - p_{s_n}(\boldsymbol{X}_0)]. \qquad (35)
\end{aligned}
$$

Using similar arguments in bounding $\mathrm{E}|\eta_4^{(1)}|$, we can show

$$\mathrm{E}|\tau(\boldsymbol{X})||p_{s_n}(\boldsymbol{X}_0) - \mathbb{I}\{\tau(\boldsymbol{X}_0 > 0)\}| = o(n^{-1/2}).$$

In view of (35), this yields $V_0 = \mathrm{E}\eta_1 + o(n^{-1/2})$. The proof is hence completed.

References

Audibert, J.-Y. and A. B. Tsybakov (2007). Fast learning rates for plug-in classifiers. *Ann. Statist. 35*(2), 608–633.

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res. 13*, 1063–1095.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *Ann. Statist. 30*(4), 927–961.

Chakraborty, B., E. B. Laber, and Y.-Q. Zhao (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials 11*(4), 408–417.

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

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Ann. Statist. 9*(3), 586–596.

Fan, C., W. Lu, R. Song, and Y. Zhou (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 79*(5), 1565–1582.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, F. Rossi, and R. Ulerich (2015). *GNU Scientific Library Reference Manual (Version 2.1)*.

Härdle, W. (1990). *Applied nonparametric regression*, Volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Liang, S., W. Lu, R. Song, and L. Wang (2017). Sparse concordance-assisted learning for optimal treatment decision. *Journal of machine learning research* (just-accepted).

Lu, W., H. H. Zhang, and D. Zeng (2013). Variable selection for optimal treatment decision. *Stat. Methods Med. Res. 22*(5), 493–504.

Luedtke, A. and A. Chambaz (2017). Faster rates for policy learning. *arXiv preprint arXiv:1704.06431*.

Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist. 44*(2), 713–742.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 331–366.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Stat. Med. 24*(10), 1455–1481.

Qian, M. and S. A. Murphy (2011). Performance guarantees for individualized treatment rules. *Ann. Statist. 39*(2), 1180–1210.

Robins, J., M. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiol. 11*, 550–560.

Steinwart, I. and A. Christmann (2008). *Support vector machines*. Information Science and Statistics. Springer, New York.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* (just-accepted).

Watkins, C. and P. Dayan (1992). Q-learning. *Mach. Learn. 8*, 279–292.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*(4), 1010–1018.

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*(3), 681–694.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics 71*(4), 895–904.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2016). Interpretable dynamic treatment regimes. *arXiv preprint arXiv:1606.01472*.

Zhao, T., G. Cheng, and H. Liu (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist. 44*(4), 1400–1437.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc. 107*(499), 1106–1118.

Zhao, Y.-Q., D. Zeng, E. B. Laber, and M. R. Kosorok (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc. 110*(510), 583–598.

Zhou, S., X. Shen, and D. A. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist. 26*(5), 1760–1782.

Zhu, R., D. Zeng, and M. R. Kosorok (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc. 110*(512), 1770–1784.