Received: 20 November 2019 | Accepted: 24 June 2020

DOI: 10.1111/biom.13337

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Resampling-based confidence intervals for model-free robust inference on optimal treatment regimes

Yunan Wu[1] | **Lan Wang**[2]

[1] School of Statistics, University of Minnesota, Minneapolis, Minnesota

[2] Department of Management Science, University of Miami, Coral Gables, Florida

Correspondence
Lan Wang, Department of Management Science, University of Miami, Coral Gables, FL 33146.
Email: lanwang@mbs.miami.edu

Funding information
NSF, Grant/Award Numbers: DMS-1712706, OAC-1940160, FRGMS-1952373

Abstract

We propose a new procedure for inference on optimal treatment regimes in the model-free setting, which does not require to specify an outcome regression model. Existing model-free estimators for optimal treatment regimes are usually not suitable for the purpose of inference, because they either have nonstandard asymptotic distributions or do not necessarily guarantee consistent estimation of the parameter indexing the Bayes rule due to the use of surrogate loss. We first study a smoothed robust estimator that directly targets the parameter corresponding to the Bayes decision rule for optimal treatment regimes estimation. This estimator is shown to have an asymptotic normal distribution. Furthermore, we verify that a resampling procedure provides asymptotically accurate inference for both the parameter indexing the optimal treatment regime and the optimal value function. A new algorithm is developed to calculate the proposed estimator with substantially improved speed and stability. Numerical results demonstrate the satisfactory performance of the new methods.

KEYWORDS
confidence interval, individualized treatment rule, inference, optimal treatment regime, weighted bootstrap

1 | INTRODUCTION

Applications in medicine, public policy, internet marketing, and other scientific areas often require estimating an individualized treatment rule (or regime, policy) to maximize the potential benefit. Several successful methods have been developed for estimating an optimal treatment regime, including Q-learning (Watkins and Dayan, 1992; Murphy, 2005b; Chakraborty *et al.*, 2010; Qian and Murphy, 2011; Song *et al.*, 2015), A-learning (Robins *et al.*, 2000; Murphy, 2003, 2005a; Moodie and Richardson, 2010; Shi *et al.*, 2018), model-free methods (Robins *et al.*, 2008; Orellana *et al.*, 2010; Zhang *et al.*, 2012; Zhao *et al.*, 2012, 2015; Athey and Wager, 2017; Linn *et al.*, 2017; Zhou *et al.*, 2017; Zhu *et al.*, 2017; Lou *et al.*, 2018; Qi *et al.*, 2018; Wang *et al.*, 2018), tree or list-based methods (Laber and Zhao, 2015; Cui *et al.*, 2017; Zhu *et al.*, 2017; Zhang *et al.*, 2018), targeted learning ensembles approach (Díaz *et al.*, 2018), among others.

This paper focuses on inference for optimal treatment regimes. In practice, it is often desirable to have an interpretable treatment regime. Here, we focus on the popular class of index rules, given by $\mathbb{D} = \{\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0) : \boldsymbol{\beta} \in \mathbb{B}\}$, where $\mathrm{I}(\cdot)$ is the indicator function and $\mathbb{B}$ is a compact subset of $\mathbb{R}^p$. We consider two important inference targets: one is the parameter $\boldsymbol{\beta}_0$ indexing the theoretically optimal treatment regime and the other is the theoretically optimal value function $V(\boldsymbol{\beta}_0)$. The former inference problem helps understand the importance of different predictors on making an optimal decision, while the latter aims to quantify the maximally achievable expected performance that can be used as a gold standard to evaluate alternative treatment regimes.

Although there exists a rich literature on estimation, the associated inference problem has not been studied until recently. For Q-learning, several inference methods have been investigated. Laber *et al.* (2014) proposed a novel locally consistent adaptive confidence interval for $\boldsymbol{\beta}_0$, Chakraborty *et al.* (2013) proposed a practically convenient adaptive m-out-of-n bootstrap for inference on $\boldsymbol{\beta}_0$, Chakraborty *et al.* (2014) introduced a double bootstrap approach for inference for $V(\boldsymbol{\beta}_0)$, Song *et al.* (2015) considered inference for $\boldsymbol{\beta}_0$ based on the asymptotic distribution theory for penalized Q-learning. Recently, Jeng *et al.* (2018) developed Lasso-based procedure for inference on $\boldsymbol{\beta}_0$ in the A-learning framework. However, accurate inference based on Q-learning and A-learning needs reliable model specification. Luedtke and Van Der Laan (2016) developed interesting theory for inference for $V(\boldsymbol{\beta}_0)$ under exceptional laws. Their approach requires to estimate the conditional treatment effect either based on a working model or in a completely nonparametric fashion.

Different from current state-of-the-art methods that are mostly model-based, we aim to develop a model-free approach for making inference for both $\boldsymbol{\beta}_0$ and $V(\boldsymbol{\beta}_0)$. This would be useful to alleviate the sensitivity of inference with respect to the underlying generative model, the specification of which is often challenging in real data analysis. It is known that the parameter indexing the optimal treatment regime $\boldsymbol{\beta}_0$ corresponds to the parameter of the Bayes rule of a weighted classification problem (Qian and Murphy, 2011; Zhang *et al.*, 2012; Zhao *et al.*, 2012). A substantial challenge in inference for $\boldsymbol{\beta}_0$ lies in the nonsmoothness of the decision function. A popular approach is to replaces the 0-1 loss by a computationally convenient surrogate loss such as the hinge loss (Zhao *et al.*, 2012; Zhou *et al.*, 2017; Lou *et al.*, 2018) or the logistic loss (Jiang *et al.*, 2019). However, existing theory (eg, Fisher consistency, generalization error bound) that justifies the use of the surrogate loss is usually derived when the form of the decision rule is unconstrained and approximated in a reproducible kernel Hilbert space. There is no guarantee that when we consider the class of decision rules $\mathbb{D}$, use of surrogate loss still leads to a decision function whose sign matches $\text{sign}(\boldsymbol{x}^T\boldsymbol{\beta}_0)$, see Lin (2002). On the other hand, robust estimator (Zhang *et al.*, 2012) that directly estimates the Bayes rule has a cubic root convergence rate and a nonnormal limiting distribution, as recently revealed in Wang *et al.* (2018). Inference is challenging due to the nonstandard asymptotics as naive bootstrap procedure is not consistent. Goldberg *et al.* (2014) proposed a SoftMax Q-learning approach to alleviate the nonsmoothness problem in Q-learning but have not explore the associated inference theory.

This paper first proposes a smoothed model-free estimator for the optimal treatment regime and introduce a proximal algorithm that substantially improves both the computational speed and the accuracy. We prove that the smoothed robust estimator has an asymptotic normal distribution and converges to $\boldsymbol{\beta}_0$ with a rate that can be made arbitrarily close to $n^{-1/2}$. We then rigorously justify the validity of a resampling approach for inference.

The remaining of the paper is organized as follows. Section 2 introduces the new method and algorithm. Section 3 carefully studies the statistical properties for estimation and inference. Section 4 reports the results from Monte Carlo simulations. Section 5 analyzes a clinical data set from the Childhood Adenotonsillectomy Trial (CHAT). Section 6 concludes with some discussions. The Appendix summarizes the technical assumptions. The online supplemental file contains the proofs and additional numerical results.

2 | PROPOSED METHODS

2.1 | Problem setup

Let A be a binary variable (0 or 1) denoting the treatment. For each subject, we observe a vector of covariates $\boldsymbol{x} \in \mathbb{R}^p$ and an outcome $Y \in \mathbb{R}$. Without loss of generality, we assume that larger outcome is preferred. To evaluate the treatment effect, we adopt the potential or counterfactual outcome framework (Rubin, 1978; Neyman, 1990) for causal inference. Let Y_1^* and Y_0^* be the potential outcome had the subject received treatment 1 and 0, respectively. In reality, we observe either Y_1^* or Y_0^*, but never both. It is assumed that the observed outcome is the potential outcome corresponding to the treatment the subject actually receives (consistency assumption in causal inference), that is, $Y = Y_1^*A + Y_0^*(1-A)$. Assume A and $\{Y_0^*, Y_1^*\}$ are independent conditional on $\boldsymbol{x}$, that is, no unmeasured confounding. In addition, we assume that the stable unit treatment value assumption (Rubin, 1986) and the positivity assumption are both satisfied, where the former requires a subject's outcome from receiving a treatment is not influenced by the treatment received by other subjects and the latter requires that $0 < P(A = a|\boldsymbol{x}) < 1, \forall\, \boldsymbol{x}$, almost surely.

An individualized treatment rule or a treatment regime, denoted by $d(\boldsymbol{x})$, is a mapping from the space of covariates to the set of treatment options $\{0, 1\}$. Let $Y^*(d)$ be the potential outcome had a subject with covariates $\boldsymbol{x}$ received the treatment assigned by $d(\boldsymbol{x})$. We have $Y^*(d) = Y_1^* d(\boldsymbol{x}) + Y_0^*\{1 - d(\boldsymbol{x})\}$. Given a collection $\mathbb{D}$ of treatment regimes, the optimal regime $\arg\max_{d\in\mathbb{D}} \mathrm{E}(Y^*(d))$ leads to the maximal average outcome if being implemented in the population.

For a given $\boldsymbol{\beta} \in \mathbb{B}$, we sometimes write the corresponding treatment regime $\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0)$ as $d_{\boldsymbol{\beta}}(\boldsymbol{x})$ or $d_{\boldsymbol{\beta}}$ for simplicity. The value function $V(\boldsymbol{\beta}) = \mathrm{E}\{Y^*(d_{\boldsymbol{\beta}})\}$ measures the

effectiveness of the treatment regime d_β. We are interested in estimating the parameter indexing the optimal rule

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} V(\boldsymbol{\beta}). \tag{1}$$

For identifiability, we assume that there exists a covariate with a nonzero coefficient whose conditional distribution given the other covariates is continuous and its coefficient is normalized to have absolute value one. Without loss of generality, we assume x_1 is a predictor that satisfies the condition. We write $\boldsymbol{\beta} = (\beta_1, \tilde{\boldsymbol{\beta}}^T)^T \in \mathbb{R}^p$. Correspondingly, we write $\boldsymbol{x} = (x_1, \tilde{\boldsymbol{x}}^T)^T$. More discussions on alternative identifiability condition can be found in Section 6.2.

2.2 | Challenges of inference based on existing robust estimators

Qian and Murphy (2011), Zhang *et al.* (2012), Zhao *et al.* (2012), among other, observed that optimal treatment regime estimation can be reformulated as a weighted classification problem. Specifically, the value function $V(\boldsymbol{\beta})$ can be equivalently expressed as

$$V(\boldsymbol{\beta}) = \mathrm{E}\left[\frac{Y}{\pi(A,\boldsymbol{x})} I\{A = d_\beta(\boldsymbol{x})\}\right], \tag{2}$$

where $\pi(A,\boldsymbol{x}) = P(A = 1|\boldsymbol{x})$ is the propensity score of the treatment and is equal to 0.5 in a randomized trial. Expression (2) is the foundation for robust or policy-search estimators for optimal treatment regime, which aim to alleviate the practical difficulty of specifying a reliable generative regression model.

A robust estimator can be obtained by directly maximizing an unbiased sample estimator of the expectation in (2), which was the approach in Zhang *et al.* (2012). In a randomized trial, based on the observed data $\{(\boldsymbol{x}_i, Y_i, A_i), i = 1, \ldots, n\}$, which are independent copies of $(\boldsymbol{x}, Y, A)$, $V(\boldsymbol{\beta})$ can be consistently estimated by its sample analog

$$V_n(\boldsymbol{\beta}) = \frac{2}{n}\sum_{i=1}^{n}\{A_i \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) + (1 - A_i)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} \le 0)\}Y_i. \tag{3}$$

Leaving out the terms in $V_n(\boldsymbol{\beta})$ that do not depend on $\boldsymbol{\beta}$, we can estimate $\boldsymbol{\beta}_0$ by

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} M_n(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n}(2A_i - 1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)Y_i. \tag{4}$$

However, as revealed in Wang *et al.* (2018) such a direct estimator for the Bayes rule belongs to a class of nonstandard M estimators. It converges at a cubic-root rate to a nonnormal limiting distribution that is characterized by the maximizer of a centered Gaussian process with a parabolic drift. The nonstandard asymptotics is a consequence of the so-called *sharp-edge effect* (Kim and Pollard, 1990). Inference based on this approach is challenging due to the nonstandard asymptotics as the naive bootstrap procedure is not consistent. The smoothed estimator we propose alleviates the sharp-edge effect caused by the indicator function and leads to faster convergence rate.

2.3 | Smoothed model-free inference for optimal treatment regime

To facilitate inference, we study an alternative estimator that can be considered as a compromise between the two robust estimation approaches described in Section 2.2. For clarity of presentation, we assume that the data are collected from a randomized trial. Instead of replacing the indicator function with the hinge loss function, we replace it with a smoothed approximation. Formally, we estimate $\boldsymbol{\beta}_0$ by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n &= \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \widetilde{M}_n(\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n}(2A_i - 1)K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right)Y_i,\end{aligned} \tag{5}$$

where $K(\cdot)$ is a smoothed approximation to the indicator function, and h_n is a sequence of smoothing parameter that goes to zero as $n \to \infty$. The function $K(\cdot)$ is required to satisfy some general regularity conditions given in the Appendix, see also Remark 1 in Section 3.1.

The motivation for the above new estimator is threefold. First, as h_n goes to zero at an appropriate rate, the parameter indexing the optimal treatment regime or the Bayes rule can be estimated at a rate arbitrarily close to $n^{-1/2}$, see Section 3.1. Second, smoothing the indicator function circumvents the aforementioned nonstandard asymptotics and would lead to a feasible bootstrap inference procedure with theoretical guarantee, see Section 3.2. Third, it also alleviates the computational challenge due to nonsmoothness, see Section 2.4 for a new efficient algorithm.

For inference, we apply a resampling technique called "weighted bootstrap" that assigns independent and identically distributed positive random weights to each observation. This resampling scheme was proposed in Rubin (1981). Barbe and Bertail (1995) provided a comprehensive introduction, see also Ma and Kosorok (2005) and Cheng and Huang (2010) for recent interesting developments. The

bootstrapped estimate of the smoothed robust estimator is defined as

$$\widehat{\boldsymbol{\beta}}_n^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \widetilde{M}_n^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n} r_i(2A_i-1)K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right)Y_i, \quad (6)$$

where $r_1,\ldots,r_n$ are random weights satisfying conditions given in Section 3.2. To evaluate the distribution of $\widehat{\boldsymbol{\beta}}_n^*$ in practice, we repeatedly generate independent samples of random weights. Following notation introduced earlier, let $\widehat{\boldsymbol{\beta}}_n^* = (\widehat{\beta}_{n1}^*, \widetilde{\boldsymbol{\beta}}_n^{*T})^T$, where $|\widehat{\beta}_{n1}^*| = 1$ and $\widetilde{\boldsymbol{\beta}}_n^* = (\widehat{\beta}_{n2}^*,\ldots,\widehat{\beta}_{np}^*)^T$. For $j = 2,\ldots,p$, let $\xi_j^{*(\alpha/2)}$ and $\xi_j^{*(1-\alpha/2)}$ be the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantile of the bootstrap distribution of $(nh_n)^{1/2}(\widetilde{\boldsymbol{\beta}}_j^* - \widetilde{\boldsymbol{\beta}}_j)$, respectively, where α is a small positive number. We can estimate $\xi_j^{*(\alpha/2)}$ and $\xi_j^{*(1-\alpha/2)}$ from a large number of bootstrap samples. An asymptotic $100(1-\alpha)\%$ bootstrap confidence interval for β_{0j}, $j = 2,\ldots,p$, is given by

$$\left\{\widetilde{\boldsymbol{\beta}}_j - (nh_n)^{-1/2}\xi_j^{*(1-\alpha/2)}, \widetilde{\boldsymbol{\beta}}_j - (nh_n)^{-1/2}\xi_j^{*(\alpha/2)}\right\}. \quad (7)$$

Next, we consider inference for the optimal value. Define

$$V_n^*(\boldsymbol{\beta}) = \frac{2}{n}\sum_{i=1}^{n} r_i\{A_i\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) + (1-A_i)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} \le 0)\}Y_i. \quad (8)$$

Note that $V_n^*(\boldsymbol{\beta})$ can be considered as a perturbed version of the V_n defined in (3). Let $d^{*(\alpha/2)}$ and $d^{*(1-\alpha/2)}$ be the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantile of the bootstrap distribution of $n^{1/2}\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\}$, respectively. An asymptotic $100(1-\alpha)\%$ bootstrap confidence interval for $V(\boldsymbol{\beta}_0)$ is

$$\left\{V_n(\widehat{\boldsymbol{\beta}}_n) - n^{-1/2}d^{*(1-\alpha/2)}, V_n(\widehat{\boldsymbol{\beta}}_n) - n^{-1/2}d^{*(\alpha/2)}\right\}. \quad (9)$$

2.4 | A proximal algorithm

The smoothed robust estimator largely alleviates the computational challenge due to the nonsmooth indicator function. However, the objective function is still a nonconvex function of the parameter. Such nonconvexity is inherent to robust estimation of optimal treatment regime (Qian and Murphy, 2011). We employ a proximal gradient descent algorithm, originally proposed in Nesterov (2007), which applies to a large class of nonconvex problems. In our setting, this algorithm substantially improves the computational speed and can accommodate high-dimensional covariates.

Consider an optimization problem with an objective function $\Phi(\boldsymbol{\beta})$. Nesterov (2007) assumes that $\Phi(\boldsymbol{\beta})$ has the decomposition $\Phi(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Psi(\boldsymbol{\beta})$, over a convex set Q, where f is a differentiable function but not necessarily convex, and Ψ is closed and convex on Q. In our setting, we take $-\widetilde{M}_n(\boldsymbol{\beta})$ as the f function, and set $\Psi(\boldsymbol{\beta}) \equiv 0$. Following Nesterov (2007), we generate a sequence of iterates $\{\boldsymbol{\beta}^{(t)}, t = 0,1,2,\ldots\}$ such that

$$\begin{aligned}\boldsymbol{\beta}^{(t)} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{B}}\{-\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}) - \langle\nabla\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t-1)}\rangle \\ &\quad + \alpha_t\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t-1)}\|^2 + \Psi(\boldsymbol{\beta})\} \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{B}}\left\{-\frac{2}{n}\sum_{i=1}^{n}(2A_i-1)K'\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)}}{h_n}\right)\right. \\ &\quad \left.\times\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{t-1})}{h_n}Y_i + \alpha_t\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t-1)}\right\|^2\right\},\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product between two vectors. Observe that the above minimization problem has a closed-form solution

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} + (n\alpha_t)^{-1}\sum_{i=1}^{n}(2A_i-1)K'\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)}}{h_n}\right)\frac{\boldsymbol{x}_i}{h_n}Y_i.$$

Hence the algorithm can be updated efficiently. The algorithm stops when $\widetilde{M}_n(\boldsymbol{\beta}^{(t)}) < \widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}) + \langle\nabla\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}), \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\rangle - \alpha_t||\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}||^2$, where α_t is a sequence of small positive numbers. To choose α_t, inspired by Fan *et al.* (2018), we employ an expanding series, which ensures that the stepsize diminishes during the update process. Details for this algorithm is provided in the supplementary material.

It is worth emphasizing that this algorithm can be easily adapted to the high-dimensional setting by taking $\Psi(\boldsymbol{\beta})$ as a regularization function, such as the L_1 penalty function.

3 | STATISTICAL PROPERTIES

3.1 | Consistency and asymptotic normality of the smoothed estimator

To lay the foundation for inference, we first present the statistical properties of the smoothed robust estimator $\widehat{\boldsymbol{\beta}}_n$ defined in (5). All the regularity conditions are summarized in the Appendix. Theorem 1 shows that $\widehat{\boldsymbol{\beta}}_n$ is consistent for the parameter indexing the optimal treatment regime. Comparing with the asymptotic normality result in Theorem 2, the consistency requires very mild

conditions and serves as a precursor step for proving asymptotic normality.

Theorem 1. *Under (A1)-(A3) and assume $K(\cdot)$ satisfies (K1), then $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + o_p(1)$.*

Recall that for identification, we write $\boldsymbol{\beta}_0 = (\beta_{01}, \widetilde{\boldsymbol{\beta}}_0^T)^T \in \mathbb{R}^p$ where $|\beta_{01}| = 1$. Similarly, we write $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1}, \widetilde{\boldsymbol{\beta}}_n^T)^T \in \mathbb{R}^p$ where $|\widehat{\beta}_{n1}| = 1$. With the above consistency result, we have $P(\widehat{\beta}_{n1} = \beta_{01}) \to 1$ as $n \to \infty$. In the following, we focus on studying the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_n$. To this end, we introduce some additional notations. Define $S(z, \widetilde{\boldsymbol{x}}) = \mathrm{E}(Y_1^* - Y_0^* | z, \widetilde{\boldsymbol{x}})$, where $z = \boldsymbol{x}^T\boldsymbol{\beta}_0$. Note that there is a one-to-one transformation between $(z, \widetilde{\boldsymbol{x}})$ and $\boldsymbol{x} = (x_1, \widetilde{\boldsymbol{x}}^T)^T$. Hence, $S(z, \widetilde{\boldsymbol{x}})$ is a measure of the conditional treatment effect. Let $S^{(1)}(z, \widetilde{\boldsymbol{x}})$ denote the partial derivative of $S(z, \widetilde{\boldsymbol{x}})$ with respect to z. Furthermore, we define

$$\boldsymbol{D} = a_1 \mathrm{E}\left\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) \mathrm{E}(Y_1^{*2} + Y_0^{*2} | z = 0, \widetilde{\boldsymbol{x}})\right\}, \quad (10)$$

$$\boldsymbol{Q} = a_2 \mathrm{E}\left\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) S^{(1)}(0, \widetilde{\boldsymbol{x}})\right\}, \quad (11)$$

where $f(z|\widetilde{\boldsymbol{x}})$ denotes the conditional probability density function of z given $\widetilde{\boldsymbol{x}}$, $a_1 = 2\int\{K'(v)\}^2 dv$, and $a_2 = \int vK''(v)dv$, with $K'(\cdot)$ and $K''(\cdot)$ denoting the first- and second-derivative of $K(\cdot)$, respectively. Note that $\boldsymbol{D}$ and $\boldsymbol{Q}$ both depend on unknown functions, for example, $f(z|\widetilde{\boldsymbol{x}})$, and are complex to approximate analytically. This motivates us to consider a bootstrap approach for inference procedure.

Theorem 2. *Assume $K(\cdot)$ satisfies (K1)-(K3) for some $b \geq 2$, $h_n = o(n^{-1/(2b+1)})$ and $n^{-1}h_n^{-4} = o(1)$. Then under (A1)-(A5),*

(1) $\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0) \to N(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-1})$ in distribution as $n \to \infty$.

(2) $\sqrt{n}\{V_n(\widehat{\boldsymbol{\beta}}_n) - V(\boldsymbol{\beta}_0)\} \to N(0, U)$ in distribution as $n \to \infty$, where $V_n(\cdot)$ is defined in (3) and $U = Var\{Y^*(d_{\boldsymbol{\beta}_0})\} + E\{(Y^*(d_{\boldsymbol{\beta}_0})^2\}$.

Remark 1. Theorem 2 implies that $\widetilde{\boldsymbol{\beta}}_n$ achieves a convergence rate arbitrarily close to $n^{-b/(2b+1)}$. The cumulative distribution function of $N(0, 1)$ satisfies these regularity conditions with $b = 2$, and would produce a convergence rate arbitrarily close to $n^{-2/5}$. With a carefully designed $K(\cdot)$ function that satisfied (K1)-(K3) with b sufficiently large, the convergence rate can be further improved. For example, $K(v) = [0.5 + \frac{105}{64}\{\frac{v}{5} - \frac{5}{3}(\frac{v}{5})^3 + \frac{7}{5}(\frac{v}{5})^5 - \frac{3}{7}(\frac{v}{5})^7\}]I(-5 \leq v \leq 5) + I(v > 5)$ satisfies (K1)-(K3) with $b = 4$. This choice leads to a convergence rate of $n^{-4/9}$. This function first appeared in Horowitz (1992), which dealt with smoothing estimator in a different setting. Our setting and proofs are very different. Especially, our proofs substantially simplified the traditional methods for handling a smoothed objective function. Example 2 in Section S7 of the supplementary material demonstrates that the performance of the smoothed estimator is not sensitive to the choice of $K(\cdot)$ in finite samples. We would recommend the distribution function of N(0,1) as the default choice due to its simplicity, which we observe to have satisfactory performance in a variety of settings.

Remark 2. The key components of the proofs are modern empirical process techniques. In particular, we introduce some recent empirical process results (Giné and Sang, 2010; Mason, 2012) on VC classes of functions that involve smoothing parameters, which were originally developed for uniform asymptotics with data-driven bandwidth selection and have not been applied to the types of problems considered here. These new techniques lead to simpler proof and are of independent interest. Our technical derivation for this and other results in the paper employ recent techniques developed by Giné and Sang (2010) and Mason (2012) for VC classes of functions that involve smoothing parameters, see the Appendix. Carefully handling function classes involving a smoothing parameter is nontrivial. The literature usually either impose a lower positive bound on h to avoid the process to blow up or requires more involved computation on the entropy bound for such classes. In contrast, the new techniques are based on a geometric argument and avoid the usually intensive entropy computation. The asymptotic normality result in part (2) of the theorem is mostly due to the fact that the estimated value function $V_n(\boldsymbol{\beta})$ is a sample average of functions that enjoy the Donsker property. Furthermore, the population value function $V(\boldsymbol{\beta})$ has gradient zero at the true value $\boldsymbol{\beta}_0$.

3.2 | Justification for resampling-based inference

Let $r_1, \ldots, r_n$ be a random sample from a distribution of a positive random variable with mean one and variance one. Assume the random weights $r_1, \ldots, r_n$ are independent of the data. Recall that

$$\widehat{\boldsymbol{\beta}}_n^* = \arg\max_{\boldsymbol{\beta} \in \mathbb{B}} \widetilde{M}_n^*(\boldsymbol{\beta})$$

$$= \arg\max_{\boldsymbol{\beta} \in \mathbb{B}} \frac{2}{n} \sum_{i=1}^n r_i(2A_i - 1)K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right) Y_i.$$

Hence, two different sources of randomness contribute to the distribution of $\widehat{\boldsymbol{\beta}}_n^*$ in this setup: one due to the random data and the other due to the random weights.

We next provide a rigorous justification for the validity of the bootstrap procedures proposed in Section 2.3. We establish that the bootstrap distribution asymptotically imitates the distribution of the original estimator. Let $r = \{r_1, \ldots, r_n\}$ be the collection of the random bootstrap weights and $w = \{W_1, \ldots, W_n\}$ be the random sample of observations, where $W_i = (\boldsymbol{x}_i, A_i, Y_i)$.

Given a sequence of random variables R_n, $n = 1, \ldots, n$, we write $R_n = o_{p_r}(1)$ if for any $\epsilon > 0, \delta > 0$, we have $P_w(P_{r|w}(|R_n| > \epsilon) > \delta) \to 0$ as $n \to \infty$. In the bootstrap literature, R_n is said to converge to zero in probability, conditional on the data.

Theorem 3. *Under (A1)-(A3), (A6) and assume $K(\cdot)$ satisfies (K1), then*

(1) $\widehat{\boldsymbol{\beta}}_n^* = \widehat{\boldsymbol{\beta}}_n + o_{p_r}(1)$;
(2) $\sqrt{n}\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\} = N(0, U) + o_{p_r}(1)$.

Part (2) of Theorem 3 suggests that we can use the perturbed value function defined in (8) with the plugged-in estimator $\widehat{\boldsymbol{\beta}}_n$ to estimate the asymptotic variance of the estimated optimal value in Theorem 2. This establishes the asymptotic validity of the confidence interval in (9), which allows for inference for the value function. The validity of the confidence interval in (7) for $\boldsymbol{\beta}_0$ is ensured by Theorem 4 below.

Theorem 4. *Assume $K(\cdot)$ satisfies (K1)-(K3) for some $b \geq 2$, $h_n = o(n^{-1/(2b+1)})$, and $\log(n) = o(nh_n^4)$. Under (A1)-(A6), $\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n^* - \widetilde{\boldsymbol{\beta}}_n) = N(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-1}) + o_{p_r}(1)$.*

Remark 3. The proofs of Theorems 3 and 4 make use of the recent results which allow for using an unconditional argument to derive conditional results. The use of the unconditional argument can be particularly convenient to combine with the Donsker class properties.

To better understand the behavior of the proposed inference procedure, we also study the properties of the smoothed estimator and its bootstrapped version under a moving parameter or local asymptotic framework. See Section S4 of the online supplementary material.

4 | SIMULATION RESULTS

We generate random data from the model $Y = \exp(\boldsymbol{x}^T\boldsymbol{\eta}) + A\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim N(0,1)$, $\boldsymbol{x} = (x_0, x_1, x_2, x_3)^T = (x_0, \widetilde{\boldsymbol{x}}^T)^T$, $x_0 = 1$ and $\widetilde{\boldsymbol{x}}$ follows a 3-dimensional multivariate normal distribution with mean zero and identity covariance matrix. We set $\boldsymbol{\eta} = (-1, -0.5, 0.5, -0.5)^T$, and consider two settings for $\boldsymbol{\beta}$. In setting 1, we have $\boldsymbol{\beta} = (-2, -2, 2, 2)^T$; while in setting 2 we have $\boldsymbol{\beta} = (-2, -2, 2, 0)^T$ with x_3 being an inactive variable for the optimal treatment regime. The optimal treatment regime is given by $\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} \leq 0)$. As discussed in Section 2.1, for identifiability, we adopt the normalization $|\beta_1| = 1$, corresponding to the coefficient of the continuous covariate x_1. Under this normalization, the population parameter indexing the optimal treatment regime is $\boldsymbol{\beta}^{\text{opt}} = (\beta_0^{\text{opt}}, \beta_1^{\text{opt}}, \beta_2^{\text{opt}}, \beta_3^{\text{opt}}) = (-1, -1, 1, 1)$ in setting 1, and $(-1, -1, 1, 0)$ in setting 2. We consider 1000 simulation runs and three different sample sizes $n = 300, 500, 1000$ in the simulation experiments. The confidence intervals are constructed based on 500 bootstrap estimates for each simulation run. That is, for each simulation run, we generate 500 independent samples of size n of positive random weights from a distribution with mean one and variance one and apply them to weight the original observations according to (6).

We first study the finite sample performance of the smoothed robust estimator in Section 2.3. The smoothed robust estimator is computed using the proximal algorithm in Section 2.4, where we choose $K(\cdot)$ to be the cumulative distribution function of standard normal distribution and set $h_n = 0.9n^{-0.2}\min\{\text{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}), \text{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta})/1.34\}$, as suggested in Silverman (1986), where "std" denotes the standard deviation function, and "IQR" denotes the interquartile range. The initial estimator $\boldsymbol{\beta}^0$ in the proximal algorithm is set as $(0, \ldots, 0)^T$. We compare the smoothed estimator with three alternative estimators. The first is the nonsmoothed estimator in (4), which was computed using the genetic algorithm, using the "genoud" function in R package "rgenoud" (Mebane and Sekhon, 2011), as suggested in Zhang *et al.* (2012). The second is the estimator based on the hinge loss (Zhao *et al.*, 2012), calculated using the function *owl* in the R package DTRlearn2 (Chen *et al.*, 2019). The third is the estimator using logistic loss, calculated using the function *glmnet* in the R package glmnet (Friedman *et al.*, 2010). Table 1 reports the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value.

The results in Table 1 demonstrates that the smoothed robust estimate has smaller bias and substantially smaller standard deviation comparing with the other three estimators, particular for the smaller sample size setting. It also

TABLE 1 Monte Carlo estimates of the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value

n	Method	β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	Match ratio	$V_n(\widehat{\beta}_n)$
Setting 1							
300	Smooth	−0.05 (0.30)	0 (0)	0.01 (0.27)	0.04 (0.31)	99.35%	−0.02 (0.17)
	Nonsmooth	−0.29 (1.45)	0.00 (0.09)	0.12 (1.21)	0.24 (1.43)	96.67%	0.06 (0.17)
	Hinge	−0.46 (0.41)	0 (0)	0.04 (0.27)	−0.04 (0.29)	91.85%	−0.05 (0.18)
	Logistic	−0.46 (0.47)	0 (0)	0.06 (0.42)	0.26 (0.57)	94.17%	−0.02 (0.18)
500	Smooth	−0.01 (0.19)	0 (0)	0.01 (0.20)	0.02 (0.22)	99.73%	0.00 (0.13)
	Nonsmooth	−0.15 (0.41)	0 (0)	0.06 (0.36)	0.13 (0.42)	98.19%	0.05 (0.13)
	Hinge	−0.37 (0.30)	0 (0)	0.01 (0.18)	−0.06 (0.20)	92.93%	−0.03 (0.13)
	Logistic	−0.41 (0.29)	0 (0)	0.04 (0.30)	0.23 (0.36)	94.61%	−0.01 (0.13)
1000	Smooth	−0.01 (0.14)	0 (0)	0.00 (0.13)	0.01 (0.15)	99.88%	−0.01 (0.09)
	Nonsmooth	−0.07 (0.24)	0 (0)	0.02 (0.22)	0.06 (0.25)	99.04%	0.03 (0.09)
	Hinge	−0.36 (0.24)	0 (0)	0.01 (0.13)	−0.07 (0.14)	92.95%	−0.04 (0.09)
	Logistic	−0.38 (0.19)	0 (0)	0.02 (0.19)	0.18 (0.23)	94.61%	−0.02 (0.09)
Setting 2							
300	Smooth	0.04 (0.26)	0 (0)	0.02 (0.24)	0.02 (0.18)	99.35%	−0.01 (0.15)
	Nonsmooth	−0.26 (0.76)	0.00 (0.06)	0.11 (0.71)	0.11 (0.37)	95.78%	0.07 (0.15)
	Hinge	−3.33 (79.42)	0 (0)	0.01 (0.22)	−0.09 (0.16)	76.19%	−0.06 (0.16)
	Logistic	−0.67 (5.13)	0.00 (0.06)	0.18 (3.33)	0.23 (2.96)	90.20%	−0.02 (0.16)
500	Smooth	0.02 (0.19)	0 (0)	0.02 (0.18)	0.00 (0.13)	99.65%	−0.01 (0.11)
	Nonsmooth	−0.16 (0.52)	0 (0)	0.06 (0.42)	0.06 (0.24)	97.37%	0.05 (0.11)
	Hinge	−0.64 (1.11)	0 (0)	0.02 (0.16)	−0.10 (0.12)	88.59%	−0.07 (0.12)
	Logistic	−0.43 (0.29)	0 (0)	0.03 (0.30)	0.12 (0.20)	92.08%	−0.03 (0.12)
1000	Smooth	−0.01 (0.14)	0 (0)	0.01 (0.13)	0.00 (0.09)	99.79%	−0.01 (0.08)
	Nonsmooth	−0.08 (0.21)	0 (0)	0.03 (0.22)	0.04 (0.17)	98.55%	0.03 (0.08)
	Hinge	−0.56 (0.24)	0 (0)	0.01 (0.12)	−0.10 (0.08)	89.69%	−0.06 (0.09)
	Logistic	−0.43 (0.20)	0 (0)	0.03 (0.20)	0.11 (0.15)	92.13%	−0.03 (0.09)

leads to higher match ratio. Estimators using hinge loss and logistic loss are even not consistent when the sample size increases. For $n = 300$, we observe that in one or two of the 100 simulation runs the nonsmooth estimator converges to the negative of the true value of β_1^{opt} (ie, the algorithm converges to 1 when the true value is −1), which causes the nonzero variance. This is probably due to the fact nonsmooth estimation is less stable when the sample size is relatively small. In addition, the expected value functions with the true parameter $\boldsymbol{\beta}^{\text{opt}}$ and random policy are simulated via Monte Carlo simulation with 10^7 replicates; for Setting 1, the optimal value turns out to be 1.14, and the value function with random policy is −0.47; and for Setting 2, the true optimal value is 0.93, and the value function with random policy is −0.29. When taking the computation time into consideration, the nonsmoothed estimator requires about 4 s for each run, while the smoothed estimator only needs 0.002 s. This suggests a substantial reduction in computational costs.

We next investigate the bootstrap confidence interval in Section 2.3. We construct 95% bootstrap confidence intervals for the parameters indexing the optimal treatment regime. Table 2 summarizes the empirical coverage probabilities and average interval lengths. We observe that the coverage probabilities are above 92.2% for sample sizes 500 and 1000, and above 91% for sample size 300. Despite the slight under coverage, the lengths of the confidence intervals are reasonable. As sample size increases, the length of the confidence interval decreases significantly. Accurate finite-sample coverage is harder to achieve due to the model-free, nonparametric nature of our approach. See similar observations in simulations focusing on nonregularity settings for dynamic treatment regimes, for instance, Laber *et al.* (2014) and Chakraborty *et al.* (2013). As for computation time, on average one bootstrap run takes less than 0.2 s.

Finally, we explore several nonregular settings, where the optimal treatment regimes may be nonunique,

TABLE 2 Empirical coverage probabilities and average interval lengths of the 95% bootstrap confidence intervals for $\boldsymbol{\beta}^{\text{opt}}$

n		β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}
Setting 1					
300	Coverage rate	92.6%	100%	93.2%	91.0%
	Average length	1.36	0	1.26	1.38
500	Coverage rate	92.2%	100%	93.0%	92.6%
	Average length	0.81	0	0.79	0.84
1000	Coverage rate	92.6%	100%	94.0%	93.4%
	Average length	0.54	0	0.53	0.56
Setting 2					
300	Coverage rate	93.4%	100%	92.6%	95.8%
	Average length	1.12	0	1.01	0.71
500	Coverage rate	94.2%	100%	93.8%	94.6%
	Average length	0.75	0	0.72	0.51
1000	Coverage rate	94.0%	100%	93.0%	95.4%
	Average length	0.50	0	0.48	0.35

motivated by Laber *et al.* (2014). In these cases, the parameter indexing the optimal treatment regime is not uniquely identifiable but inference for the optimal value may still be feasible. We focus here on the bootstrap confidence interval for the optimal value. In setting 3, the same data generative model as before is used with $\boldsymbol{\beta} = (1, 2, 0.02, 0)^T$. For setting 4 and 5, $\boldsymbol{\beta} = (-1, 1, 0, 0)^T$, however, the first random covariate x_1 is generated from the discrete uniform distribution on the set $\{-1, 0, 1, 2\}$ and $\{1, 2\}$, respectively, instead of the standard normal distribution. For completeness, the bootstrap confidence intervals for the optimal value in setting 1 and setting 2 are also studied.

Let p denote the probability of generating a covariate vector $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{\beta} = 0$. This is a useful measure of the nonregularity of the model (Laber *et al.*, 2014). According to this measurement, setting 1-3 are regular (R) cases with $p = 0$; while setting 4 and 5 are nonregular (NR) with $p = 0.25$ for setting 4 and $p = 0.5$ for setting 5.

Table 3 summarizes the empirical coverage rate and average length for the 95% bootstrap confidence intervals for the optimal value functions. The results demonstrate that the bootstrap confidence intervals for the optimal value have desirable coverage rates with reasonable interval lengths, even in the nonregular cases. For comparison, we also report the percentage of times these bootstrap confidence would cover the value function from a random policy. The percentage is really low, which implies that the proposed method performs much better than random assignment even in the nonregular cases.

TABLE 3 Empirical coverage probabilities and average interval lengths of the 95% confidence intervals for $V(\boldsymbol{\beta}^{\text{opt}})$

n	Setting type	1 R	2 R	3 R	4 NR	5 NR
300	Coverage rate	93.0%	92.6%	96.4%	97.2%	95.4%
	Average length	0.67	0.61	0.78	0.40	0.41
	CR for random policy	0%	0%	0%	0%	31.2%
500	Coverage rate	93.8%	94.0%	96.0%	95.2%	94.4%
	Average length	0.52	0.47	0.62	0.31	0.31
	CR for random policy	0%	0%	0%	0%	12.4%
1000	Coverage rate	93.6%	95.4%	97.0%	96.0%	96.0%
	Average length	0.37	0.33	0.43	0.22	0.22
	CR for random policy	0%	0%	0%	0%	0.8%

5 | A REAL DATA EXAMPLE

We analyze a clinical data set from the Childhood Adenotonsillectomy Trial (CHAT). This is a randomized study designed to test whether early adenotonsillectomy (eAT, denoted as treatment 1) is helpful to improve neurocognitive functioning, behavior, and quality of life for children with mild to moderate obstructive sleep apnea, compared with watchful waiting plus supportive care (WWSC, denoted as treatment 0), see Marcus *et al.* (2013). In this trial, 464 children with mild to moderate obstructive sleep apnea syndrome, ages 5-9.9 years, were randomly assigned to eAT and WWSC. Some biochemical and neurocognitive test results were recorded before the treatment and 7 months after the treatment.

We consider the baseline Apnea-Hypopnea Index (AHI), with a natural log-transformation as recommended by Marcus *et al.* (2013), as an explanatory variable. AHI is the number of apneas or hypopneas recorded during the study per hour of sleep. It is an important measurement of the quality of sleep and is commonly used by doctors to classify the severity of sleep apnea. Marcus *et al.* (2013) suggested that black children tend to experience different improvements with eAT comparing with children from other races. We hence include race (binary, 1=African American, 0 for others) as another covariate. For the outcome variable, to balance the benefits and adverse effects from eAT, we adopt a composite score. The composite score uses the ratio of the follow-up AHI and baseline AHI (both with natural log-transformations) as an effective measure of benefit. On the other hand, it takes into account the adverse events documented according to the CHAT study manual of procedures as penalty.

We estimate the optimal treatment regime in the class of treatment regimes $\mathbb{D} = \{I(\beta_0 + \beta_1\text{AHI} + \beta_2\text{race} > 0) :$

$|\beta_1| = 1\}$. The kernel function $K(\cdot)$ and the bandwidth selection are the same as in Section 4. The smoothed estimator for the baseline AHI is normalized to 1, the race is 0.56, with $(0.34, 0.97)$ as the 95% bootstrap confidence interval, and the intercept is 0.39, with confidence interval $(0.22, 0.65)$. The confidence intervals suggest that the coefficients are all significantly different from 0. The analysis suggests that it is reasonable to assign WWSC to those children with milder symptoms (lower AHI). It also suggests that black children display more improvement in the AHI scale with eAT. The results are consistent with those observed empirically in Redline *et al.* (2011), Marcus *et al.* (2013), and Dean *et al.* (2016). The average outcome with randomized treatment is 0.288. The estimated average outcome corresponding to the estimated optimal treatment regime is 0.063, with a 95% bootstrap confidence interval $(-0.126, 0.260)$. This suggested a significant reduction of the composite outcome score when applying the optimal treatment regime. To compare with the smoothed estimator, we also calculate the nonsmoothed estimator, whose coefficients are 1 for baseline AHI, -0.19 for the race, and -0.40 for the intercept. Its estimated optimal value is -0.034. The nonsmoothed estimators are significantly different from the smoothed ones. In Example 4 of Section S7 in the supplementary, we demonstrate based on fivefold cross-validation that for this real data example, the nonsmoothed estimator is quite unstable.

6 | DISCUSSIONS

6.1 | Extension to other settings

The method we propose can be extended to observational studies using the inverse probability weighting approach. Assume the propensity score $\pi(x) = P(A = 1|\boldsymbol{x})$ can be modeled as $\pi(x, \xi)$ where ξ is a finite-dimensional parameter (eg, via logistic regression). Let $\widehat{\xi}$ be an estimate of ξ. Under the commonly adopted assumption of no unmeasured confounding, a smoothed robust estimator for $\boldsymbol{\beta}_0$ can be constructed as

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1} \sum_{i=1}^{n} \frac{\left[A_i K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right) + (1-A_i)\left\{1 - K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right)\right\}\right] Y_i}{A_i\pi(x,\widehat{\xi}) + (1-A_i)(1-\pi(x,\widehat{\xi}))}. \tag{12}$$

Example 3 in Section S7 of the supplementary material confirms that this smoothed estimator provides accurate estimation for the optimal treatment regime when the propensity score model is correctly specified. The estimator in (12) can also be extended to be doubly robust similarly as in Zhang *et al.* (2012). Due to the presence of nuisance parameter, the theory of asymptotic normality and inference is more technically involved. This will be a future research topic.

It is worth pointing out that our method is applicable to binary response, as binary random variable is sub-Gaussian after centering. Example 1 in Section S7 of the supplementary materiel demonstrates that our estimation and inference procedures work effectively for binary responses. For survival outcome under random censoring, our method can be extended to obtain a robust procedure for estimating the optimal treatment regime maximizing the restricted mean survival time, similarly as in Zhao *et al.* (2015). Let $\widetilde{T}$ denote the survival time. Let $T = \min\{\widetilde{T}, \tau\}$ be the outcome of interest, where τ is the time till the end of the study. Let C denote the censoring time and $\Delta = I(T < C)$ be the censoring indicator. We observe $Y = \min\{T, C\}$. Based on the observed data $\{Y_i, \boldsymbol{x}_i, \Delta_i, A_i\}$, $i = 1, \ldots, n$ from a randomized trial, the smoothed estimator can be constructed as

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n} \sum_{i=1}^{n} \frac{\left[A_i K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right) + (1-A_i)\left\{1 - K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\right)\right\}\right]}{\widehat{G}_C(Y_i|\boldsymbol{x}, A)} \Delta_i Y_i,$$

where $G_C(t|\boldsymbol{x}, A) = P(C > t|\boldsymbol{x}, A)$ is the conditional survival function of the censoring time C given $(\boldsymbol{X}, A)$, and $\widehat{G}_C(\cdot|\boldsymbol{x}, A)$ is an estimator of $G_C(\cdot|\boldsymbol{x}, A)$.

6.2 | On the identifiability condition

The asymptotic normality results can be established under alternative identifiability constraint such as the requirement that the L_1/L_2 norm of $\boldsymbol{\beta}$ is 1, or identifiability of $\boldsymbol{\beta}$ up to a scale. However, this usually leads to more technically involved proof as $\boldsymbol{\beta}$ is constrained to be the boundary point of a unit sphere and $V(\boldsymbol{\beta})$ does not have a derivative at $\boldsymbol{\beta}$. This issue was often ignored in the theory development in many existing literature, which only adjust for the constraint in an ad hoc way in the numerical implementation. See Zhu and Xue (2006) for more discussions in an index model setting and a careful delete-one-component method to handle this rigorously.

For identifiability, we assume that there exists a covariate whose conditional distribution given the other covariates is absolutely continuous. This is a common assumption for index model and is satisfied in many real applications. In practice, domain experts may help suggest such a candidate continuous covariate and the statisticians can run confirmatory analysis (eg, comparing the conditional treatment effect conditional on this covariate) to verify if this is a viable choice. In the case when all

relevant covarites are discrete (eg, gender and race), the problem reduces to comparing a finite number of decision rules and the main target of inference is arguably the optimal value. Our simulation settings 4 and 5 only include discrete variables in the optimal regime. The simulation results in Table 3 show that our proposed bootstrap confidence interval still provides reasonable empirical coverage probability for the optimal value in discrete cases.

6.3 | Nonregular settings

The optimal treatment regime may not be unique if there exists a subpopulation who responds similarly to the two treatment options. In such a setting, the complexity of nonregularity arises, see the discussions in Robins (2004), Moodie and Richardson (2010), Laber *et al.* (2014), Song *et al.* (2015), and Luedtke and Van Der Laan (2016). Uniform inference under nonregularity or exceptional laws is a challenging problem.

Although our theory does not apply to this scenario, our simulation results show that our bootstrap confidence interval for the optimal value function displays a fair degree of robustness in the two examples where nonregularity occurs. As an example, in simulation setting 5, if $x_1 = 1$, then the subject responds the same to the two treatment options; while if $x_1 = 2$, the subject benefits from treatment 1. There are four decision rules of interest for this example. The optimal treatment rule is nonunique as one may assign either treatment 0 (say no treatment or a standard, less expensive treatment) or treatment 1 to those subjects with $x_1 = 1$. A relative simple approach to breaking the nonuniqueness is to introduce a secondary criterion. For example, one may argue that under the principle of avoiding over-treatment, there exists a unique optimal decision rule of interest, in this case $I(x_1 = 2)$, which would not assign treatment 1 when ambiguity exists in order to reduce costs and avoid potential risks. Based on the sample, this unique optimal treatment regime can be consistently estimated by selecting the decision rule that maximizes the sample average treatment effect while treating the smallest proportion of the population.

There are additional inference targets that have rarely been discussed in the literature, that is, inference about the linear combination in the rule $\boldsymbol{x}^T\boldsymbol{\beta}$ or about the rule itself $I(\boldsymbol{x}^T\boldsymbol{\beta} > 0)$. These two quantities are of interest in clinical practice as they indicate how much confidence we can put on the prescribed optimal decision. We are currently studying these inference problems and will report the results in a future article.

ACKNOWLEDGMENTS

The authors thank the co-editor, the AE, and two anonymous referees for their constructive comments that have helped us significantly improve the paper. The authors acknowledge financial support from NSF DMS-1712706, NSF OAC-1940160, and NSF FRGMS-1952373.

ORCID

Lan Wang https://orcid.org/0000-0002-3217-0202

REFERENCES

Apostol, T.M. (1974) *Mathematical Analysis*, Vol. 2. Reading, MA: Addison-Wesley.

Athey, S. and Wager, S. (2017). *Efficient Policy Learning*. Stanford, CA: Institute for Economic Policy Research.

Barbe, P. and Bertail, P. (1995) *The Weighted Bootstrap*. New York, NY: Springer.

Chakraborty, B., Laber, E.B. and Zhao, Y. (2013) Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69, 714–723.

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

Chakraborty, B., Murphy, S. and Strecher, V. (2010) Inference for nonregular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19, 317–343.

Chen, Y., Liu, Y., Zeng, D. and Wang, Y. (2019) *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*. R package version 1.0.

Cheng, G. and Huang, J.Z. (2010) Bootstrap consistency for general semiparametric m-estimation. *Annals of Statistics*, 38, 2884–2915.

Cui, Y., Zhu, R. and Kosorok, M. (2017) Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic Journal of Statistics*, 11, 3927–3953.

Dean, II, D.A., Goldberger, A.L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., Sahoo, S.S., Jayapandian, C.P., Cui, L., Morrical, M.G., Surovec, S., Zhang, G.Q. and Redline, S. (2016) Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*, 39, 1151–1164.

Díaz, I., Savenkov, O. and Ballman, K. (2018) Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika*, 105, 723–738.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018) I-lamm for sparse learning: simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics*, 46, 814–841.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Giné, E. and Sang, H. (2010) Uniform asymptotics for kernel density estimators with variable bandwidths. *Journal of Nonparametric Statistics*, 22, 773–795.

Goldberg, Y., Song, R., Zeng, D. and Kosorok, M.R. (2014) Comment on "dynamic treatment regimes: technical challenges and applications". *Electronic Journal of Statistics*, 8, 1290.

Horowitz, J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica*, 60, 505–531.

Jeng, X.J., Lu, W. and Peng, H. (2018) High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12, 2074–2089.

Jiang, B., Song, R., Li, J. and Zeng, D. (2019) Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29, 1633–1710.

Kim, J.K. and Pollard, D. (1990) Cube root asymptotics. *Annals of Statistics*, 18, 191–219.

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

Laber, E. and Zhao, Y. (2015) Tree-based methods for individualized treatment regimes. *Biometrika*, 102, 501–514.

Lin, Y. (2002) Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6, 259–275.

Linn, K.A., Laber, E.B. and Stefanski, L.A. (2017) Interactive q-learning for probabilities and quantiles. *Journal of the American Statistical Association*, 112, 638–649.

Lou, Z., Shao, J. and Yu, M. (2018) Optimal treatment assignment to maximize expected outcome with multiple treatments. *Biometrics*, 74, 506–516.

Luedtke, A.R. and Van Der Laan, M.J. (2016) Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44, 713.

Ma, S. and Kosorok, M.R. (2005) Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96, 190–217.

Marcus, C.L., Moore, R.H., Rosen, C.L., Giordani, B., Garetz, S.L., Taylor, H.G., *et al.* (2013) A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368, 2366–2376.

Mason, D.M. (2012) Proving consistency of non-standard kernel estimators. *Statistical Inference for Stochastic Processes*, 15, 151–176.

Mebane, W.R., Jr, and Sekhon, J.S., (2011) Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, 42, 1–26.

Moodie, E.E. and Richardson, T.S. (2010) Estimating optimal dynamic regimes: correcting bias under the null. *Scandinavian Journal of Statistics*, 37, 126–146.

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

Murphy, S.A. (2005a) An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24, 1455–1481.

Murphy, S.A. (2005b) A generalization error for q-learning. *Journal of Machine Learning Research*, 6, 1073–1097.

Nesterov, Y. (2007) Gradient methods for minimizing composite objective function. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Neyman, J. D. M. Dabrowska and T. P. Speed, (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5, 465–472.

Orellana, L., Rotnitzky, A.G. and Robins, J. (2010) Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part ii: proofs of results. *The International Journal of Biostatistics*, 6, 9.

Qi, Z. and Liu, Y. (2018) D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, 12, 3601–3638.

Qian, M. and Murphy, S.A. (2011) Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39, 1180–1210.

Redline, S., Amin, R., Beebe, D., Chervin, R.D., Garetz, S.L., Giordani, B., et al. (2011) The childhood adenotonsillectomy trial (chat): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. *Sleep*, 34, 1509–1517.

Robins, J.M. (2004) *Optimal Structural Nested Models for Optimal Sequential Decisions*. New York, NY: Springer.

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

Robins, J., Orellana, L. and Rotnitzky, A. (2008) Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27, 4678–4721.

Rubin, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6, 34–58.

Rubin, D.B. (1981) The Bayesian bootstrap. *Annals of Statistics*, 9, 130–134.

Rubin, D.B. (1986) Which ifs have causal answers. *Journal of the American Statistical Association*, 81, 961–962.

Shi, C., Fan, A., Song, R. and Lu, W. (2018) High-dimensional *a*-learning for optimal dynamic treatment regimes. *Annals of Statistics*, 46, 925–957.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. New York, NY: Chapman and Hall.

Song, R., Wang, W., Zeng, D. and Kosorok, M. (2015) Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25, 901–920.

Wang, L., Zhou, Y., Song, R. and Sherwood, B. (2018) Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113, 1243–1254.

Watkins, C.J. and Dayan, P. (1992) Q-learning. *Machine Learning*, 8, 279–292.

Zhang, Y., Laber, E.B., Davidian, M. and Tsiatis, A.A. (2018) Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113, 1541–1549.

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

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

Zhao, Y.-Q., Zeng, D., Laber, E.B., Song, R., Yuan, M. and Kosorok, M.R. (2015) Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102, 151–168.

Zhao, Y., Zeng, D., Rush, A.J. and Kosorok, M.R. (2012) Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107, 1106–1118.

Zhou, X., Mayer-Hamblett, N., Khan, U. and Kosorok, M.R. (2017) Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112, 169–187.

Zhu, L. and Xue, L. (2006) Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B*, 68, 549–570.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S. and Zhao, H. (2017) Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73, 391–400.

SUPPORTING INFORMATION

Web Appendices that contain the proofs, additional numerical results (referenced in Sections 2.4, 3.2, 5 and 6) and the R code to reproduce the simulations are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Wu Y, Wang L. Resampling-based confidence intervals for model-free robust inference on optimal treatment regimes. *Biometrics*. 2021;77:465–476. https://doi.org/10.1111/biom.13337

APPENDIX: REGULARITY CONDITIONS

We first state some regularity conditions, where (K1)-(K3) are assumptions imposed on $K(\cdot)$, while (A1)-(A6) are assumptions imposed on the data.

(K1) $K(\cdot)$ is twice differentiable, $K(\cdot)$, $K'(\cdot)$, and $K''(\cdot)$ all bounded variation on the real line. Furthermore, $\lim_{\nu\to-\infty} K(\nu) = 0$, $\lim_{\nu\to\infty} K(\nu) = 1$; $\int\{K'(\nu)\}^2 d\nu$ and $\int\{K''(\nu)\}^2 d\nu$ are both finite.

(K2) For some integer $b \geq 2$, and any $1 \leq i \leq b$, $\int |\nu^i K'(\nu)| d\nu < \infty$; $\int_{-\infty}^{\infty} \nu^i K'(\nu) d\nu = 0$ for $1 \leq i \leq b-1$ and $\int_{-\infty}^{\infty} \nu^b K'(\nu) d\nu = d \neq 0$.

(K3) For any integer i between 0 and b, any $\eta > 0$, and any sequence $\{h_n\}$ converging to 0, $\lim_{n\to\infty} h_n^{i-b} \int_{|h_n\upsilon|>\eta} |\nu^i K'(\nu)| d\nu = 0$, and $\lim_{n\to\infty} h_n^{-1} \int_{|h_n\nu|>\eta} |K''(\nu)| d\nu = 0$.

(A1) $\mu(a, \boldsymbol{x})$ is bounded for almost all $\boldsymbol{x}$, and $a = 0, 1$; $Y_a^* - \mu(a, \boldsymbol{x})$, $a = 0, 1$, has a sub-Gaussian distribution for almost every $\boldsymbol{x}$.

(A2) The support of the distribution of $\boldsymbol{x}$ is not contained in any proper linear subspace of $\mathbb{R}^p$. For almost every $\tilde{\boldsymbol{x}}$, the distribution of x_1 conditional on $\tilde{\boldsymbol{x}}$ has everywhere a positive density. The components of $\tilde{\boldsymbol{x}}$ are bounded by M_x.

(A3) Let $S(z, \tilde{\boldsymbol{x}}) = \mathrm{E}\{Y_1^* - Y_0^* | z, \tilde{\boldsymbol{x}}\}$, where $z = \boldsymbol{x}^T\boldsymbol{\beta}_0$. For almost every $\tilde{\boldsymbol{x}}$, $S(0, \tilde{\boldsymbol{x}}) = 0$. And for every $\epsilon > 0$, $\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\epsilon} \mathrm{E}\{\mathrm{I}(x^T\boldsymbol{\beta} > 0) S(z, \tilde{\boldsymbol{x}}) f(z|\tilde{\boldsymbol{x}})\} < \mathrm{E}\{\mathrm{I}(x^T\boldsymbol{\beta}_0 > 0) S(z, \tilde{\boldsymbol{x}}) f(z|\tilde{\boldsymbol{x}})\}$.

(A4) Given any integer $0 \leq i \leq b-1$, for all z in a neighborhood of 0, $f^{(i)}(z|\tilde{\boldsymbol{x}})$ is a continuous function of z and satisfies $|f^{(i)}(z|\tilde{\boldsymbol{x}})| < M_f$ for almost every $\tilde{\boldsymbol{x}}$, where $M_f > 0$ is a constant.

(A5) Let $S^{(i)}(0, \tilde{\boldsymbol{x}})$, $i = 0, 1, \ldots, b$, denote the ith partial derivative of $S(z, \tilde{\boldsymbol{x}})$ with respect to z. For $0 \leq i \leq b$, for all z in a neighborhood of 0, $S^{(i)}(z, \tilde{\boldsymbol{x}})$ is a continuous function of z and satisfies $|S^{(i)}(z, \tilde{\boldsymbol{x}})| < M_S$ for almost every $\tilde{\boldsymbol{x}}$, where $M_s > 0$ is a constant. The matrices $\mathrm{E}\{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T f(0|\tilde{\boldsymbol{x}}) S^{(1)}(0, \tilde{\boldsymbol{x}})\}$ and $-\mathrm{E}\{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{\beta}}_0) f(0|\tilde{\boldsymbol{x}}) S^{(1)}(0, \tilde{\boldsymbol{x}})\}$ are negative definite.

(A6) The random weights $r_1, \ldots, r_n$ form a random sample from a distribution of a positive random variable with mean one and variance one. Assume that $r_i - \mathrm{E}(r_i)$ has a sub-Gaussian distribution, $i = 1, \ldots, n$.

Remark 4. The bounded variation assumption on $K(\cdot)$, $K'(\cdot)$, and $K''(\cdot)$ are relatively weak (Apostol (1974, chapter 6)). This and other assumptions in (K1)-(K2) are satisfied if $K(\cdot)$ is taken to be the distribution function of standard normal distribution ($b = 2$) or the function in Remark 1 ($b = 4$). However, $K(\cdot)$ is not required to be a cumulative distribution function. The bounded variation assumption implies that $K(\cdot)$, $|K'(\cdot)|$, and $|K''(\cdot)|$ are uniformly bounded. Our assumptions on the data are also relatively mild. Condition (A1) imposes mild assumption on the tail distribution of $Y_a^* - \mu(a, \boldsymbol{x})$, $a = 0, 1$, and allows for both normal distribution and many other nonnormal distributions. Condition (A3) is a margin type condition to ensure identification of $\boldsymbol{\beta}_0$.