Received: 25 November 2020 | Accepted: 12 May 2022

DOI: 10.1111/biom.13719

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Nonparametric inverse-probability-weighted estimators based on the highly adaptive lasso

Ashkan Ertefaie[1] | Nima S. Hejazi[2] | Mark J. van der Laan[3]

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, USA

[2]Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine

[3]Division of Biostatistics, School of Public Health and Department of Statistics, University of California, Berkeley, California, USA

Correspondence
Ashkan Ertefaie, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA.
Email: ashkan_ertefaie@urmc.rochester.edu

Funding information
National Institute on Alcohol Abuse and Alcoholism, Grant/Award Number: R21AA027571; National Institute of Neurological Disorders and Stroke, Grant/Award Number: R61NS120240; National Institute on Drug Abuse, Grant/Award Number: R01DA048764

Abstract

Inverse-probability-weighted estimators are the oldest and potentially most commonly used class of procedures for the estimation of causal effects. By adjusting for selection biases via a weighting mechanism, these procedures estimate an effect of interest by constructing a pseudopopulation in which selection biases are eliminated. Despite their ease of use, these estimators require the correct specification of a model for the weighting mechanism, are known to be inefficient, and suffer from the curse of dimensionality. We propose a class of nonparametric inverse-probability-weighted estimators in which the weighting mechanism is estimated via undersmoothing of the highly adaptive lasso, a nonparametric regression function proven to converge at nearly $n^{-1/3}$-rate to the true weighting mechanism. We demonstrate that our estimators are asymptotically linear with variance converging to the nonparametric efficiency bound. Unlike doubly robust estimators, our procedures require neither derivation of the efficient influence function nor specification of the conditional outcome model. Our theoretical developments have broad implications for the construction of efficient inverse-probability-weighted estimators in large statistical models and a variety of problem settings. We assess the practical performance of our estimators in simulation studies and demonstrate use of our proposed methodology with data from a large-scale epidemiologic study.

KEYWORDS
adaptive estimation, causal inference, efficient influence function, semiparametric efficiency

1 | INTRODUCTION

Inverse-probability-weighted (IPW) estimators have been widely used in a diversity of fields, as inverse probability weighting allows for the adjustment of selection biases by the assignment of weights (i.e., based on propensity scores) to observational units such that a pseudopopulation mimicking the target population is generated. The construction of IPW estimators is relatively straightforward, as the only nuisance parameter that must be estimated is the propensity score. Owing in part to the ease with which IPW estimators may be constructed, their application has been frequent in causal inference (e.g., Robins et al., 2000), missing data (e.g., Robins et al., 1994), and survival analysis (e.g., Tsiatis, 2007).

While inverse probability weighting may easily be implemented and is appropriate for use in a variety of problem settings, the resultant estimators face several disadvantages. Unfortunately, IPW estimators require a correctly specified estimate of the propensity score to produce consistent estimates of the target parameter and can be inefficient in certain settings (e.g., randomized controlled trials). What is more, IPW estimators suffer from the curse of dimensionality, as their rate of convergence depends entirely on the convergence rate of the postulated model for the propensity score. This latter requirement has proven a significant obstacle to investigators wishing to use data-adaptive techniques in the estimation of propensity scores. To overcome these significant shortcomings, van der Laan (2014) proposed the *targeted* IPW estimator, which facilities the use of data-adaptive techniques in estimating the relevant weight functions. While the targeted estimator is asymptotically linear, it has been shown to suffer from issues of irregularity (van der Laan, 2014). Alternatively, doubly robust estimation procedures, which are based on constructing models for both the propensity score and the outcome mechanism (Bang & Robins, 2005), were proposed. Doubly robust estimators are consistent for the target parameter when either one of the two nuisance parameters is consistently estimated; moreover, such estimators are efficient when both nuisance parameter estimators are correctly specified (Rotnitzky et al., 1998; van der Laan and Robins, 2003). While doubly robust estimators allow two opportunities for consistent estimation, their performance depends critically on the choice of estimators of these nuisance parameters. When finite-dimensional models are used to estimate the two nuisance parameters, doubly robust estimators may perform poorly, due to the possibility of model misspecification in either of the nuisance parameter estimators (Cao et al., 2009; Kang & Schafer, 2007; Vermeulen & Vansteelandt, 2015, 2016). Although doubly robust procedures facilitate the use of data-adaptive techniques for modeling nuisance parameters, the resultant estimator can be irregular with large bias and a slow rate of convergence when either of the nuisance parameters is inconsistently estimated. To ease such issues, van der Laan (2014) proposed a targeted doubly robust estimator that does not suffer from the irregularity issue; the properties of this estimation procedure were investigated in detail by Benkeser et al. (2017).

Though many data-adaptive regression techniques have been shown to provide consistent estimates in flexible models, establishing the rate of convergence for such approaches is often a significant obstacle. Among such approaches, the highly adaptive lasso (HAL) stands out for its ability to flexibly estimate arbitrary functional forms with a fast rate of convergence under relatively mild conditions. The HAL is a nonparametric regression function that minimizes a loss-specific empirical risk over linear combinations of indicator basis functions under the constraint that the sum of the absolute value of the coefficients is bounded by a constant (van der Laan, 2017; van der Laan & Bibaut, 2017). Letting the space of the functional parameter be a subset of the set of càdlàg (right-hand continuous with left-hand limits) functions with sectional variation norm bounded by a finite constant, van der Laan (2017) showed that the HAL estimator converges to the true function at a rate faster than $n^{-1/4}$, regardless of dimensionality d (i.e., for any fixed d). Bibaut and van der Laan (2019) subsequently improved this convergence rate to $n^{-1/3}\log(n)^{d/2}$. Unlike most existing data-adaptive techniques that require local smoothness assumptions on the true functional form, the finite sectional variation norm assumption imposed by HAL constitutes a (less restrictive) global smoothness assumption, making it a powerful approach for use in a variety of settings; in Section S1 of the Supporting Information, we briefly compare the sectional variation norm to another commonly used approach.

Hirano et al. (2003) proposed an efficient IPW estimator in which the propensity score is estimated in a sieve approach by the logit series estimators (Geman & Hwang, 1982). Their approach has two shortcomings: (1) it requires that both propensity score and outcome models be continuously differentiable, with the level of smoothness of the propensity score increasing (by factor of 7) with the covariate dimension (Hirano et al., 2003, Assumption 4(i)); and (2) the rate of convergence of the estimated propensity score depends on the covariate dimension. We overcome these issues by proposing the first dimension-free and smoothness-free efficient nonparametric IPW estimator.

We show that IPW estimators can be asymptotically (nonparametric) efficient when the propensity score is estimated using an HAL estimator tuned in a particular manner. Specifically, we show that undersmoothing of the HAL estimator allows for the resultant IPW estimator of the target parameter to be asymptotically linear and a solution to an appropriate efficient influence function (EIF) equation. In the typical construction of HAL estimators, cross-validation is used to determine the sectional variation norm of the underlying functional parameter. By contrast, undersmoothing of the HAL estimator selects a sectional variation norm greater than the choice made by the (global) cross-validation selector. A significant challenge arises in finding a suitable choice of sectional variation norm—one that results in sufficient undersmoothing while simultaneously avoiding overfitting. We provide theoretical conditions under which the desired degree of undersmoothing may be achieved, and we supplement our theoretical investigations by providing practical guidance for appropriately choosing the required tuning parameters. Our proposed approach obviates many of the challenges associated with current methods of choice, namely,

(i) in contrast with standard IPW estimators, our estimators do not suffer from an *asymptotic* curse of dimensionality, allowing asymptotic efficiency;
(ii) in contrast with targeted IPW estimators, our estimators do not suffer from potential issues of irregularity;

(iii) in contrast with doubly robust estimators, our IPW estimators rely on only a single nuisance parameter and may be formulated without the EIF; and
(iv) in contrast with the method of Hirano et al. (2003), our estimators do not require local smoothness assumptions on the propensity score or the outcome models.

2 | PRELIMINARIES

2.1 | Problem formulation, notation, and target parameter

Consider data generated by typical cohort sampling: let $O = (W, A, Y) \sim P_0 \in \mathcal{M}$ be the data on a given observational unit, where P_0, the distribution of O, lies in the nonparametric model $\mathcal{M}$. The random variable $W \in \mathcal{W}$ constitutes baseline covariates measured prior to treatment $A \in \{0, 1\}$, and Y is an outcome of interest. Suppose we observe a sample of n independent and identically distributed units $O_1, \ldots O_n$, whose empirical distribution we denote by P_n. We let $Pf = \int f(o)dP(o)$ for a given function $f(o)$ and distribution P, denoting by $\mathbb{E}_P$ expectations with respect to P. Let $G : \mathcal{M} \to \mathcal{G}$ be a functional nuisance parameter where $\mathcal{G} = \{G(P) : P \in \mathcal{M}\}$. We use $G := G(P) \equiv G(P)(A \mid W)$ to denote the treatment mechanism under an arbitrary distribution $P \in \mathcal{M}$. We refer to the treatment mechanism under the true data-generating distribution as G_0, that is, $G_0 := G(P_0)$. Letting Y^a be the potential outcome that would have been observed under the intervention that sets A to level a, we define the full (unobservable) data unit X as $X = \{W, Y^0, Y^1\} \sim P_X \in \mathcal{M}^F$. A common parameter of interest is the mean counterfactual outcome under treatment, that is, $\Psi^F(P_X) = \mathbb{E}_{P_X}(Y^1)$, where $\Psi^F : \mathcal{M}^F \to \mathbb{R}$ and $\mathcal{M}^F$ is the nonparametric model for the full data X. While we present our results in the context of this target parameter, we stress that our developments extend to any arbitrary $a \in \mathcal{A}$ without loss of generality. Define the corresponding full data canonical gradient $D^F(X, \Psi^F) = \{Y^1 - \Psi^F(P_X)\}$ and allow $\prod$ to be a projection operator in the Hilbert space $L_0^2(P)$ with inner product $\langle h_1 h_2 \rangle = \mathbb{E}_P(h_1 h_2)$. To identify the causal effect of interest, we assume consistency (i.e., $Y_i = Y^{A_i}$) and no unmeasured confounding (i.e., $A \perp Y^a \mid W$) (Hernán & Robins, 2020). We also make the strong positivity assumption that, for all $W \in \mathcal{W}$, $\min(G_0, 1 - G_0) > \delta$ where δ is a positive constant. Consistency links the potential outcomes to those observed, no unmeasured confounding is a particular case of the randomization (i.e., coarsening at random) assumption, and positivity ensures that there is sufficient treatment assignment variation for the treatment effect to be assessed.

2.2 | Inverse-probability-weighted mapping

We define an IPW mapping of $D^F(X, \Psi^F)$ so as to estimate the target parameter $\Psi(P_X)$ using the observed data:

$$U_G(O; \Psi) = \frac{AY}{G(1 \mid W)} - \Psi(P),$$

where $\Psi : \mathcal{M} \to \mathbb{R}$. Here, $\Psi(P) = \mathbb{E}_P(Y^1) = \mathbb{E}_P\{\mathbb{E}_P(Y \mid A = 1, W)\}$. Under the standard identification assumptions noted in Section 2.1, $\mathbb{E}_P\{U_G(O; \Psi) \mid X\} = D^F(X, \Psi^F)$. Under coarsening at random, the tangent space of G may be defined as $T_{\text{CAR}} = \{\eta(A, W) : \mathbb{E}_{P_0}\{\eta(A, W) \mid W\} = 0\}$. The canonical gradient of Ψ at a distribution $P \in \mathcal{M}$ is

$$D^\star(P) = U_G(O; \Psi) - D_{\text{CAR}}(P),$$

where $D_{\text{CAR}}(P) = \prod\{U_G(\Psi) \mid T_{\text{CAR}}\}$ (Robins et al., 1994; van der Laan and Robins, 2003). Following van der Laan and Robins (2003), we have that $\prod\{U_G(\Psi) \mid T_{\text{CAR}}\} = \mathbb{E}_P\{U_G(O; \Psi) \mid A = 1, W\} - \mathbb{E}_P\{U_G(O; \Psi) \mid W\}$, which may equivalently be expressed

$$D_{\text{CAR}}(P) = \frac{A - G(A \mid W)}{G(A \mid W)} Q(1, W),$$

where $Q(1, W) = \mathbb{E}_P(Y \mid A = 1, W)$ is the conditional mean outcome.

2.3 | The highly adaptive lasso estimator

The HAL is a nonparametric regression function with the capability to estimate infinite-dimensional functional parameters at a fast rate (roughly $n^{-1/3}$) (van der Laan, 2017; van der Laan & Bibaut, 2017). Benkeser and van der Laan (2016) first demonstrated the utility of the HAL estimator in extensive simulation experiments. The zeroth-order HAL estimator constructs a linear combination of indicator basis functions to minimize the expected value of a loss function while constraining the L_1-norm of its coefficients to be bounded by a finite constant corresponding to the sectional variation norm.

Let $\mathbb{D}[0, \tau]$ be the Banach space of d-variate real-valued càdlàg functions on a cube $[0, \tau] \in \mathbb{R}^d$, where τ is the upper bound of all supports and is assumed to be finite. For each function $f \in \mathbb{D}[0, \tau]$, define the supremum norm as $\|f\|_\infty = \sup_{w \in [0,\tau]} |f(w)|$. The d-dimensional cube $[0, \tau]$ can be represented as a union of lower dimensional cubes (i.e., l-dimensional with $l \leq d$) and the origin. That is, $[0, \tau] = \{\cup_s (0, \tau_s]\} \cup \{0\}$ where $\cup_s$ is over all subsets s of $\{1, 2, \ldots, d\}$. For a given subset $s \subset \{0, 1, \ldots, d\}$ and for each function $f \in \mathbb{D}[0, \tau]$, we define the sth section of f as

$f_s(u) = f(u_1 I(1 \in s), \dots, u_d \in I(d \in s))$. This is the function that varies along the variables in u_s according to f while setting other variables to zero. Then, the sectional variation norm of a given f may be defined as

$$\|f\|_\nu^\star := |f(0)| + \sum_{s \subset \{1,\dots,d\}} \int_{0_s}^{\tau_s} |df_s(u)|,$$

where the sum is over all subsets of the coordinates $\{0, 1, \dots, d\}$. The term $\int_{0_s}^{\tau_s} |df_s(u)|$ is the s-specific variation norm.

Under the assumption that our nuisance functional parameter $G \in \mathbb{D}[0, \tau]$ has finite sectional variation norm, logit G may be represented as (Gill et al., 1995):

$$\begin{aligned} \text{logit}\, G(\mathrm{w}) &= \text{logit}\, G(0) + \sum_{\mathrm{s} \subset \{1,\dots,\mathrm{d}\}} \int_{0_\mathrm{s}}^{\mathrm{w}_\mathrm{s}} \text{dlogit}\, G_\mathrm{s}(\mathrm{u}) \\ &= \text{logit}\, G(0) + \sum_{\mathrm{s} \subset \{1,\dots,\mathrm{d}\}} \int_{0_\mathrm{s}}^{\tau_\mathrm{s}} \mathrm{I}(\mathrm{u}_\mathrm{s} \leq \mathrm{w}_\mathrm{s}) \text{dlogit}\, G_\mathrm{s}(\mathrm{u}). \end{aligned} \quad (1)$$

The representation in Equation (1) may be approximated using a discrete measure that places mass on each observed $W_{s,i}$, denoted by $\beta_{s,i}$. Letting $\phi_{s,i}(w_s) = I(w_s \geq u_{s,i})$, where $w_{s,i}$ are support points of logitG_s, we have

$$\text{logitG}_\beta = \beta_0 + \sum_{\mathrm{s} \subset \{1,\dots,\mathrm{d}\}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \beta_{\mathrm{s,i}} \phi_{\mathrm{s,i}},$$

where $|\beta_0| + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^n |\beta_{s,i}|$ is an approximation of the sectional variation norm of logitG. HAL first expands the covariate dimension by embedding observations in a space of up to $n(2^d - 1)$ indicator basis functions (i.e., the HAL basis features), where $(2^d - 1)$ corresponds to all subsets of the set $\{1, 2, \dots, d\}$, excluding the empty set. Let Φ denote the constructed $n(2^d - 1) \times n$ design matrix. Then, we can write $\text{logitG}_\beta = \beta_0 + \Phi^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $n(2^d - 1) \times 1$ vector of parameters. The loss-based HAL estimator $\boldsymbol{\beta}_{n,\lambda}$ is defined as

$$\boldsymbol{\beta}_{n,\lambda} = \underset{\beta : |\beta_0| + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^n |\beta_{s,i}| < \lambda}{\arg\min} P_n L(\text{logitG}_\beta),$$

where $L(\cdot)$ is an appropriate loss function and $P_n f = n^{-1} \sum_{i=1}^n f(O_i)$. Denote by $G_{n,\lambda} := G_{\beta_{n,\lambda}}$ the HAL estimate of G_0. When the functional nuisance parameter is a conditional probability (e.g., the propensity score for a binary treatment), log-likelihood loss may be used. Different choices of the tuning parameter λ result in unique HAL estimators; our goal is to select an HAL estimator that allows the construction of an asymptotically linear IPW estimator of $\Psi(P_0)$. We let λ_n denote this data adaptively selected tuning parameter. Section S1 of the Supporting Information discusses further the sectional variation norm and representation in terms of indicator basis functions.

3 | METHODOLOGY

We estimate the full data parameter $\Psi^F(P_X)$ using an IPW estimator $\Psi(P_n, G_n)$, which is a solution to the score equation $P_n U_{G_n}(\Psi) = 0$. That is,

$$\Psi(P_n, G_n) = n^{-1} \sum_{i=1}^n \frac{A_i Y_i}{G_n(A_i \mid W_i)}. \quad (2)$$

Alternatively, a stabilized IPW estimator may be defined as the solution to $n^{-1}\{A_i(Y_i - \Psi(P))\}/\{G_n(A_i \mid W_i)\} = 0$. The consistency and convergence rate of these estimators relies on the consistency and convergence rate of the estimator G_n of G_0. While finite-dimensional (i.e., parametric) models are often utilized to construct the propensity score estimator G_n, it has been widely conceded that such models are not sufficiently flexible to provide a consistent estimator of the nuisance parameter G_0. Consequently, corresponding confidence intervals for $\Psi^F(P_X)$ will have coverage tending to zero asymptotically. While flexible, data-adaptive regression techniques may be used to improve the consistency of G_n for G_0, establishing asymptotic linearity of the resultant IPW estimator $\Psi(P_n, G_n)$ can prove challenging. Specifically,

$$\begin{aligned} \Psi(P_n, G_n) - \Psi(P_0, G_0) &= P_n U_{G_n}(\Psi) - P_0 U_{G_0}(\Psi) \\ &= (P_n - P_0) U_{G_0}(\Psi) + P_0\{U_{G_n}(\Psi) - U_{G_0}(\Psi)\} \\ &\quad + (P_n - P_0)\{U_{G_n}(\Psi) - U_{G_0}(\Psi)\}. \end{aligned} \quad (3)$$

Assuming that $U_G(\Psi)$ is càdlàg with a universal bound on the sectional variation norm, it can be shown that $(P_n - P_0)\{U_{G_n}(\Psi) - U_{G_0}(\Psi)\} = o_p(n^{-1/2})$ for each G, relying only on standard empirical process theory and the assumption of consistency. Consequently, the asymptotic linearity of our IPW estimator relies on the asymptotic linearity of $P_0\{U_{G_n}(\Psi) - U_{G_0}(\Psi)\}$. As data-adaptive regression techniques have a rate of convergence slower than $n^{-1/2}$, the bias of $P_0\{U_{G_n}(\Psi) - U_{G_0}(\Psi)\}$ will dominate the right-hand side of Equation (3).

To show that asymptotic linearity of $\Psi(P_n, G_n)$ can be established when G is estimated using a properly tuned HAL estimator, we introduce Lemma 1, which is an adaptation of Theorem 1 of van der Laan et al. (2019).

Lemma 1. *Let G_{n,λ_n} be an HAL estimator of G with L_1-norm bound λ_n chosen such that*

$$\min_{(s,j)\in\mathcal{J}_n}\left\|P_n\frac{d}{d\mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}}}L(\mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}})(\phi_{\mathrm{s,j}})\right\| = \mathrm{o_p}(\mathrm{n}^{-1/2}), \tag{4}$$

where $L(\cdot)$ is log-likelihood loss and $\mathcal{J}_n$ is a set of indices for the basis functions such that $\beta_{n,s,j}\neq 0$. Let $D(f,G_n) = f\cdot(A-G_n)$. Here, f is càdlàg with finite sectional variation norm, and $\tilde{f}$ is a projection of f onto the linear span of the basis functions $\phi_{s,j}$ in $L^2(P)$, where $\phi_{s,j}$ satisfies condition (4). Assuming $\|f-\tilde{f}\|_{2,P_0} = O_p(n^{-1/4})$, it follows that $P_nD(\tilde{f},G_n) = o_p(n^{-1/2})$ and $P_nD(f,G_n) = o_p(n^{-1/2})$ where $\|f-\tilde{f}\|^2_{2,P_0} = \int(f-\tilde{f})^2(o)dP_0(o)$.

In condition (4), $d/d\mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}}\{\mathrm{L}(\mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}})(\phi_{\mathrm{s,j}})\}$ is $d/d\epsilon\{L(\mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}}+\epsilon\phi_{\mathrm{s,j}})\}$, denoting the directional derivative of the loss along the path $\mathrm{logitG}^{\epsilon}_{\mathrm{n},\lambda_\mathrm{n}} = \mathrm{logitG}_{\mathrm{n},\lambda_\mathrm{n}}+\epsilon\phi_{\mathrm{s,j}}$. Under log-likelihood loss,

$$\frac{d}{d\epsilon}L(\mathrm{logitG}^{\epsilon}_{\mathrm{n},\lambda_\mathrm{n}})(\phi_{\mathrm{s,j}})\Big|_{\epsilon=0} = (A-G_{n,\lambda_n})\phi_{s,j},$$

which is the corresponding score function. Generally, maximum likelihood estimators exactly solve the score equations; however, the L_1-norm restriction results in only approximate solutions. Condition (4) implies that the L_1-norm must be increased (i.e., weakening the restriction and undersmoothing the fit) until one of the score equations is solved to a precision of $o_p(n^{-1/2})$. Lemma 1 provides the theoretical guarantee for our main results presented in Theorem 1. The use of different loss functions leads to different estimators with different properties; however, our proposal applies to any loss function that generates scores of form $f(\boldsymbol{w})\{A-G_{n,\lambda_n}(\boldsymbol{w})\}$, including the squared error loss.

In general, undersmoothing is used for the following two reasons: (1) to produce an asymptotically efficient and unbiased estimator (plug-in or nonplug-in) of a pathwise differentiable parameter where the nuisance parameters are estimated using a data-adaptive approach; and (2) to produce an asymptotically unbiased and normally distributed estimator of a nonpathwise differentiable parameter (Wasserman, 2006). The former is our motivation in our paper and, as shown in Theorem 1, under certain assumptions, undersmoothing of the HAL estimator of the nuisance parameter G results in IPW estimators that are asymptotically linear and efficient in the nonparametric model. In the following, in a slight abuse of notation, we use f to denote $Q_0(1,W)/G_0$, which is a particular member of $\mathbb{D}[0,\tau]$ (i.e., under Assumption 1). This requires further assumptions.

Assumption 1. Let $Q_0(1,W) = \mathbb{E}(Y^1 \mid W)$ and $G_0(W)$ be càdlàg with finite sectional variation norm.

Assumption 2. Let $\tilde{f}$ be the projection of $f = Q_0(1,W)/G_0$ onto a linear span of basis functions $\phi_{s,j}$ in $L^2(P)$, for $\phi_{s,j}$ satisfying condition (4). Then, $\|f-\tilde{f}\|_{2,P_0} = O_p(n^{-1/4})$.

As the set of càdlàg functions with finite sectional variation norm contains a rich variety of functional forms, Assumption 1 is mild in that it would be expected to hold in nearly any practical application. We now consider Assumption 2. This assumption states that the degree of undersmoothing needs to be such that the generated basis functions in the HAL fit of G_n are sufficient to approximate f within an $n^{-1/4}$ neighborhood of f (i.e., $\|f-\tilde{f}\|_{2,P_0} = O_p(n^{-1/4})$). This assumption is not particularly strong. First, assuming that f is càdlàg with finite sectional variation norm (Assumption 1), there always exists λ such that $\|f-\tilde{f}\|_{2,P_0} = O_p(n^{-1/4})$ is satisfied. Second, the required convergence rate is even slower than the established rate obtained by the HAL estimator (i.e., $O_p(n^{-1/3})$). Let $f = Q_0(1,W)/G_0$ and suppose that $df_s/d\mathrm{logitG_s} < \infty$ for all sections $s\subset\{0,1,\dots,d\}$. It follows that

$$f(w) = f(0)+\sum\nolimits_{s\subset\{1,\dots,d\}}\int_{0_s}^{\tau_s}\mathbb{I}(u_s\leq w_s)df_s(u)$$

$$= f(0)+\sum\nolimits_{s\subset\{1,\dots,d\}}\int_{0_s}^{\tau_s}\mathbb{I}(u_s\leq w_s)\frac{df_s(u)}{\mathrm{d}}$$

$$\mathrm{logit}\,G_{\mathrm{s(u)}}d\mathrm{logit}\,G_{\mathrm{s}}(\mathrm{u}).$$

When $\mathrm{logit}\,G_0$ has similar complexity to f, measured by the support set for the knot points of the basis functions, Assumption 2 may hold without undersmoothing. On the other hand, when G_0 is a simple function (e.g., in randomized controlled trials), undersmoothing is more likely to be needed so that the undersmoothed $d\mathrm{logitG_n}$ has rich enough support to approximate f. In general, as f becomes more complex relative to G_0, more undersmoothing should be required. We examine this phenomenon in Section S3 of the Supporting Information. In our simulation study, we find that even in the extreme case that $G_0(W) = 0.5$ and Q_0 is a function of W, undersmoothing still improves the efficiency of HAL-based IPW estimators.

Theorem 1. *Suppose that the support of W is uniformly bounded, that is, $\mathcal{W}\subseteq[0,\tau]^d$ for some finite constant τ. Let G_{n,λ_n} be an HAL estimator of G_0 with L_1-norm bound equal to λ_n. Under Assumption 1, when λ_n is chosen such that condition (4) and Assumption 2 are satisfied, the estimator $\hat{\psi} = \Psi(P_n, G_{n,\lambda_n})$ will be asymptotically efficient with influence function*

$$\hat{\psi}-\psi_0 = P_n\{U_{G_0}(\Psi)-D_{CAR}(P_0)\}+o_p(n^{-1/2}),\quad \textit{where}\quad \psi_0 = \Psi(P_0).$$

Intuitively, Theorem 1 states that when the HAL estimator G_{n,λ_n} is properly undersmoothed, the resultant estimate will include a rich enough set of basis functions to approximate any arbitrary càdlàg function with finite sectional variation norm (as per Lemma 1). With respect to the asymptotic linearity result, Assumptions 1 and 2, along with condition (4), imply that the chosen set of basis functions must be sufficient to solve the EIF equation, that is, $P_n D_{\text{CAR}}(G_{n,\lambda_n}, Q_0) = o_p(n^{-1/2})$. A proof of this result is given in Section S2 of the Supporting Information. Conveniently, when the form of the EIF is unknown, inference is attainable via the standard bootstrap (Cai & van der Laan, 2019). The undersmoothing condition (4) provides a rate condition and cannot be used to determine the level of undersmoothing in finite sample settings. In Section 4.2, we provide practical criteria for undersmoothing that are verifiable using the observed data.

Remark 1. The undersmoothing does not impact the rate of convergence of highly adaptive estimators as long as $\|\beta_{n,\lambda}\|_{L_1}$ remains finite. The latter is plausible in our setting for two reasons: (1) we have assumed that the propensity score has finite sectional variation norm; and (2) the undersmoothing criterion is based on approximating a function with finite sectional variation norm (i.e., $Q_0(1,W)/G_0(W)$), and thus, even undersmoothing is unlikely to lead to diverging $\|\beta_{n,\lambda}\|_{L_1}$ as n increases. This is confirmed in our simulation studies presented in Figures 1 and 2 where the level of undersmoothing stabilizes as $n \to \infty$.

Remark 2. van der Laan et al. (2019) showed that when the conditional outcome model (i.e., Q_0) is estimated using an undersmoothed HAL, the resulting plug-in estimator will be efficient and asymptotically linear. Qiu et al. (2020) also explore undersmoothing methods to achieve efficiency in plug-in estimators using machine learning tools. Our results establish asymptotic efficiency of nonplug-in (i.e., IPW) estimators when propensity score (i.e., G_0) is estimated using an undersmoothed HAL. Even though an important principle in the efficiency proofs of these two types of HAL-based estimators is the same (i.e., the HAL fit must solve a large class of score equations), there are also principle differences between the two. First, using knowledge on G_0 in obtaining an HAL-based estimate of G_0 may lead to an asymptotically inefficient IPW estimator (van der Laan and Robins, 2003). For example, if it is known that A only depends on W_1, a component of W, estimating the propensity score using an undersmoothed HAL that only includes W_1 (and the corresponding derived features) may result in a highly inefficient estimator. This is because the HAL fit may not solve enough score equations to satisfy $P_n D_{\text{CAR}}(G_{n,\lambda_n}, Q_0) = o_p(n^{-1/2})$. In contrast, for a plug-in estimator any knowledge on Q_0 should be used in the formulation of HAL to enhance its efficiency. Second, in finite sample, the IPW estimators are in general more sensitive to overfitting the HAL-based estimator beyond the optimal undersmoothing level than the plug-in estimators. This is because in IPW estimators, G_0 appears in the denominator, and thus, too much undersmoothing can make the resulting estimator unstable by pushing the G_n toward zero. Hence, in finite samples, a carefully designed undersmoothing criterion is needed for IPW estimators. Finally, because the dependence of plug-in and IPW estimators on the HAL estimator is very different (e.g., IPW estimator is highly nonlinear in G_0), the proof of asymptotic linearity of these estimators requires different techniques.

4 | ESTIMATION

4.1 | Cross-fitted inverse-probability-weighted estimation

Cross-fitting has previously been studied as a step to debias parameter estimates when the relevant nuisance functions are estimated using data-adaptive techniques (Chernozhukov et al., 2017; Klaassen, 1987; Zheng & van der Laan, 2011).

To employ V-fold cross-fitting, split the data, uniformly at random, into V mutually exclusive and exhaustive sets of size approximately nV^{-1}. Denote by $P^0_{n,v}$ the empirical distribution of a training sample and by $P^1_{n,v}$ the empirical distribution of a validation sample. For a given λ, exclude a single (validation) fold of data and fit the HAL estimator using data from the remaining $(V-1)$ folds; use this model to estimate the propensity scores for observational units in the holdout (validation) fold. Repeat this process V times, such that holdout estimates of the propensity score are available for all observational units. The cross-fitted IPW estimator $\widehat{\Psi}(P^1_{n,v}, G_{n,\lambda})$ is the solution to $V^{-1}\sum_{v=1}^{V} P^1_{n,v} U_{G_{n,\lambda,v}}(\Psi) = 0$, where $G_{n,\lambda,v}(A \mid W)$ is the estimate of $G_0(A \mid W)$ applied to the training sample for the fifth sample split for a given λ.

Theorem S1, found in the Supporting Information, shows that the cross-fitted IPW estimator is asymptotically linear. Although cross-fitting relaxes the Donsker class condition to show $(P_n - P_0)\{U_{G_{n,\lambda_n}}(\Psi) - U_{G_0}(\Psi)\} = o_p(n^{-1/2})$, our proposed approach does not obviate the need for the Donsker class condition. The condition is required to show that $P_n D_{\text{CAR}}(G_{n,\lambda_n}, Q_0) = o_p(n^{-1/2})$, which requires assuming that Q_0/G_0 has finite section variation norm. Thus, cross-fitting provides only finite-sample improvements in our setting.

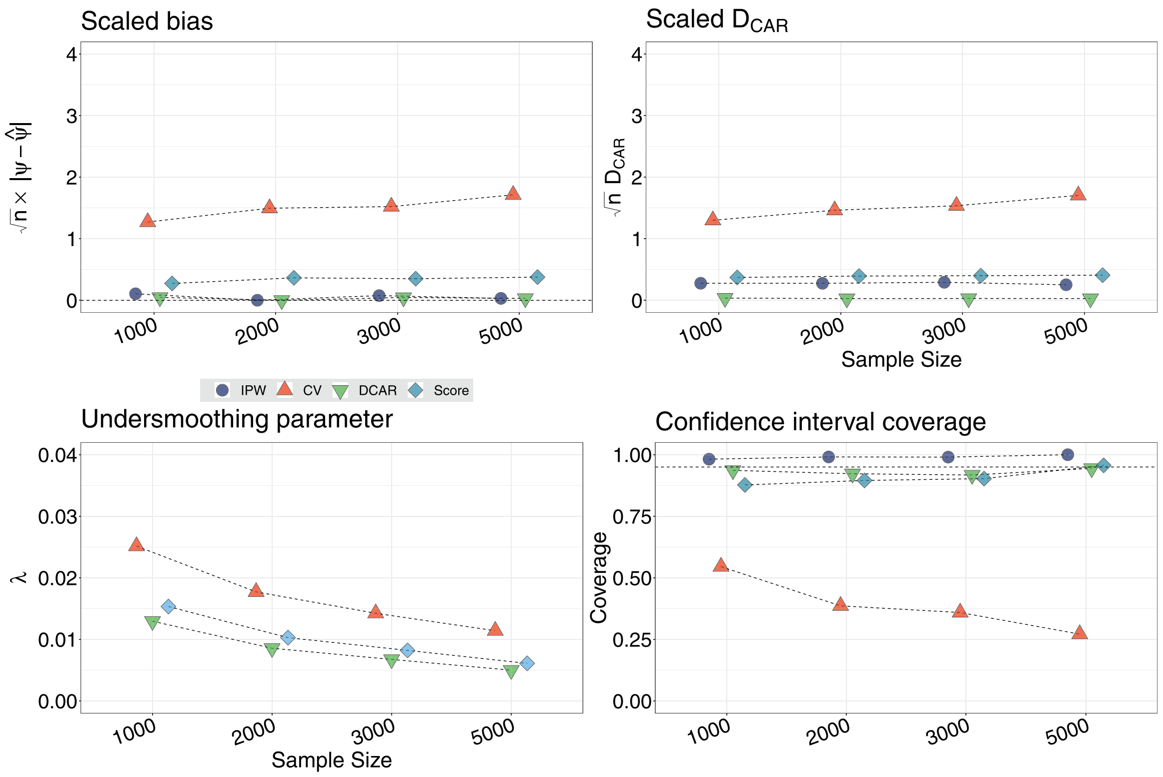

FIGURE 1 Comparative performance of IPW estimator variants in scenario 1. Circle: parametric; Triangle: nonparametric with cross-validated λ selector; "$\bigtriangledown$": D_{CAR}-based λ selector; "⋄": score-based λ selector. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

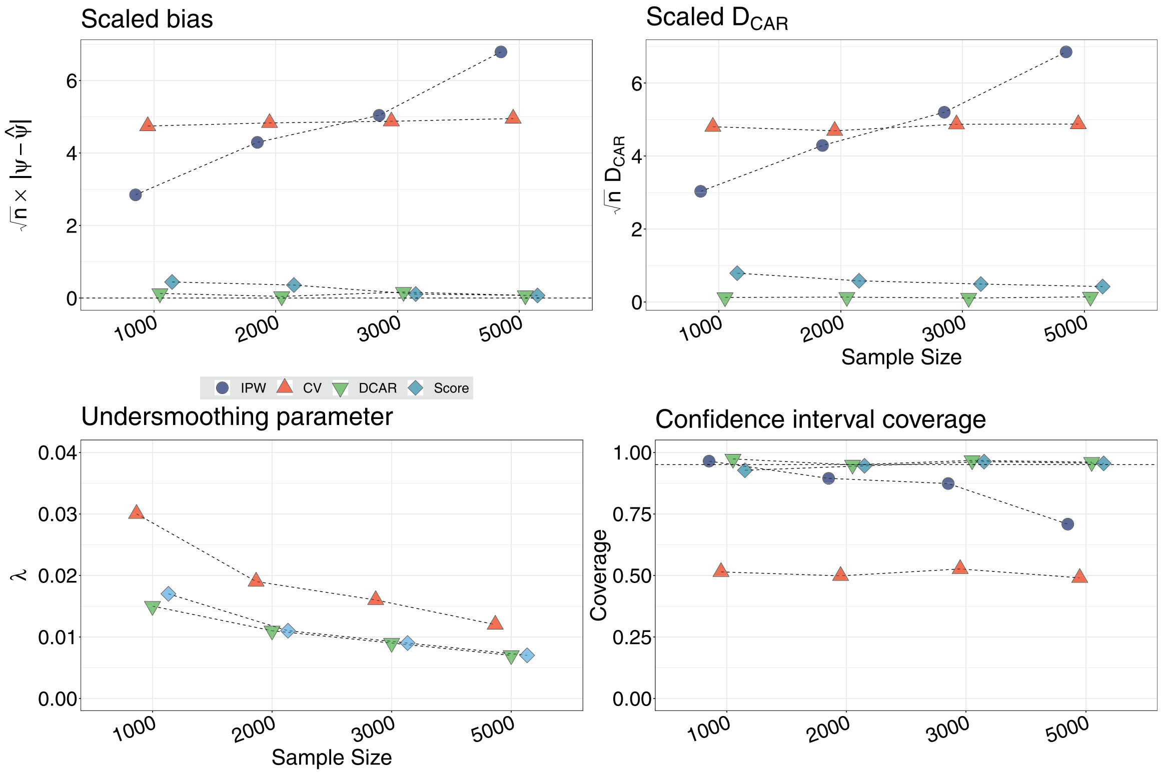

FIGURE 2 Comparative performance of IPW estimator variants in scenario 2. Circle: parametric; Triangle: nonparametric with cross-validated λ selector; "$\bigtriangledown$": D_{CAR}-based λ selector; "⋄": score-based λ selector. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

4.2 | Undersmoothing in practice

Undersmoothing is crucial for both asymptotic linearity and efficiency of our proposed estimators. Our theoretical results show that targeted undersmoothing of the HAL estimator of G can result in an IPW estimator $\hat{\psi}$ that is a solution to the EIF equation. In practice, an L_1-norm bound for an estimate of G may be obtained such that

$$\lambda_n = \arg\min_{\lambda} \left| V^{-1} \sum_{v=1}^{V} P^1_{n,v} D_{\mathrm{CAR}}(G_{n,\lambda,v}, Q_{n,v}) \right|, \qquad (5)$$

where $Q_{n,v}$ is a cross-validated HAL estimate of $Q_0(1,W)$ with the L_1-norm bound based on the global cross-validation selector. The criterion (5) is motivated by the goal of achieving efficiency asymptotically, requiring that our estimator be a solution to the EIF equation, that is, $P_n D_{\mathrm{CAR}}(G_{n,\lambda_n}, Q_0) = o_p(n^{-1/2})$.

For a general censored data problem and inverse probability of censoring weighted HAL estimator, in certain complex settings, the derivation of the EIF can become mathematically involved and tedious. This arises, for example, in longitudinal settings with many decision points. For such settings, alternative criteria that do not require knowledge of the form of the EIF may prove useful. To this end, we propose the criterion:

$$\lambda_n = \arg\min_{\lambda} V^{-1} \sum_{v=1}^{V} \left[\sum_{(s,j)\in\mathcal{J}_n} \frac{1}{\|\beta_{n,\lambda,v}\|_{L_1}} \left| P^1_{n,v} \tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) \right| \right], \qquad (6)$$

in which $\|\beta_{n,\lambda}\|_{L_1} = |\beta_{n,\lambda,0}| + \sum_{s\subset\{1,\ldots,d\}} \sum_{j=1}^{n} |\beta_{n,\lambda,s,j}|$ is the L_1-norm of the coefficients $\beta_{n,\lambda,s,j}$ in the HAL estimator $G_{n,\lambda}$ for a given λ, and $\tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) = \phi_{s,j}(W)\{A - G_{n,\lambda,v}(1 \mid W)\}\{G_{n,\lambda,v}(1 \mid W)\}^{-1}$. This score-based criterion leverages a general characteristic of canonical gradients: propensity score terms always appear in the denominator. Increasing λ results in decreasing the empirical mean of the score equation $S_{s,j}(\phi, G_{n,\lambda,v})$ (where $S_{s,j}(\phi, G_{n,\lambda,v}) = \phi_{s,j}(W)\{A - G_{n,\lambda,v}(1 \mid W)\}$) and increasing the variance of the weight function $\{G_{n,\lambda,v}(1 \mid W)\}^{-1}$. The latter occurs because, as we undersmooth the propensity score fit, the values of $G_{n,\lambda,v}$ may start approaching the boundaries of the unit interval, leading to large and unstable weights. Another key component of our score criterion is the L_1-norm $\|\beta_{n,\lambda}\|_{L_1}$. Under Assumption 1, as λ increases, the L_1-norm increases, but its rate of increase diminishes as λ diverges. Hence, at a certain point in the grid of λ, decreases in the empirical mean of $S_{s,j}(\phi, G_{n,\lambda,v})/\|\beta_{n,\lambda}\|_{L_1}$ are insufficient for satisfying condition 6, which starts increasing on account of $\{G_{n,\lambda,v}(1 \mid W)\}^{-1}$.

For both proposed undersmoothing criteria, the series of propensity score models based on HAL is constructed as follows. First, an initial fit is obtained via global cross-validation, to choose a starting value λ_{CV}. Next, undersmoothed HAL fits are constructed by weakening the restriction placed on the L_1-norm, that is, $\lambda \geq \lambda_{\mathrm{CV}}$. Then, the value of λ is increased until the target criterion is satisfied, selecting a particular HAL fit in the sequence.

Remark 3. As noted by a referee, there is no guarantee that the selected λ_n using our finite sample criteria would satisfy conditions of Theorem 1. However, the criteria correspond to the most efficient IPW estimator for a given data.

4.3 | Stability under near-violations of positivity

In practice, despite the strong positivity assumption, the estimated propensity score may fall close to the boundaries of the unit interval in finite samples. Such cases may result in large or unstable estimates of the inverse probability weights required for estimator construction. To mitigate this issue, we propose truncation of propensity score estimates. Importantly, because of the assumed positivity assumption and uniform consistency of the HAL (van der Laan & Bibaut, 2017), $\min(G_n, 1 - G_n) > \delta$ with probability 1 as $n \to \infty$. This implies that no truncation is needed asymptotically; hence, the $\sqrt{n}$–inference is not affected. For a given positive constant κ, truncation sets all propensity score estimates lower than κ or greater than $1-\kappa$ to κ and $1-\kappa$, respectively. The selectors of Equations (5) and (6) may be straightforwardly extended to achieve optimal κ-truncation:

$$(\lambda_n, \kappa) = \arg\min_{\lambda,\kappa} \left| V^{-1} \sum_{v=1}^{V} P^1_{n,v} D_{\mathrm{CAR}}(G_{n,\lambda,v,\kappa}, Q_{n,v}) \right|, \qquad (7)$$

$$(\lambda_n, \kappa) = \arg\min_{\lambda,\kappa} V^{-1} \sum_{v=1}^{V} \left[\sum_{(s,j)\in\mathcal{J}_n} \frac{1}{\|\beta_{n,\lambda,v}\|_{L_1}} \left| P^1_{n,v} \tilde{S}_{s,j}(\phi, G_{n,\lambda,v,\kappa}) \right| \right], \qquad (8)$$

where $G_{n,\lambda,v,\kappa}$ is the truncated propensity score estimate for a given λ and κ, and $\tilde{S}_{s,j}(\phi, G_{n,\lambda,v,\kappa}) = \phi_{s,j}(W)\{A - G_{n,\lambda,v,\kappa}(1 \mid W)\}\{G_{n,\lambda,v,\kappa}(1 \mid W)\}^{-1}$.

5 | NUMERICAL STUDIES

The practical performance of our proposed IPW estimators was assessed in simulation studies. We present two of these studies in the sequel, with four additional scenarios discussed in Section S3 of the Supporting Information. In the present two scenarios, we assess the performance of our IPW estimators against alternatives based on correctly specified parametric models for the propensity score and a cross-validated HAL estimator, illustrating that estimators based on undersmoothing of HAL can be made both unbiased and efficient.

In both of the following scenarios, $W_1 \sim$ Uniform $(-2, 2)$, $W_2 \sim$ Normal$(\mu = 0, \sigma = 0.5)$, $\epsilon \sim$ Normal$(\mu = 0, \sigma = 0.1)$, and $expit(x) = \{1 + \exp(-x)\}^{-1}$. In each setting, we sample $n \in \{1000, 2000, 3000, 5000\}$ independent and identically distributed observations, applying each estimator to the resultant data. This was repeated 200 times. In both scenarios, the true propensity score G_0 is bounded away from zero (i.e., $0.15 < G_0$); thus, the positivity assumption holds. In both scenarios, the true treatment effect is zero. Note that while these scenarios include two covariates and up to several thousand observations, they are far from low-dimensional with regard to the complexity of the HAL—in fact, the basis function expansion is expected to generate between 3000 (for $n = 1000$) and 15,000 indicator bases (for $n = 5000$) to represent the relevant main and interaction terms across the several thousand observational units.

In the first scenario, $A \mid W \sim$ Bernoulli$\{expit(0.75W_1 + 0.5W_2)\}$ and $Y \mid A, W = 0.5W_1 - 2/3W_2 + \epsilon$. As both models are linear, parametric IPW estimators are expected to be unbiased. In the second scenario, $A \mid W \sim$ Bernoulli$\{expit(0.5W_2^2 - 0.5\exp(W_1/2))\}$ and $Y \mid A, W = 2W_1 - 2W_2^2 + W_2 + W_1W_2 + 0.5 + \epsilon$. Due to nonlinearity of the propensity score model, the parametric IPW estimator (with main effect terms) is expected to exhibit bias, whereas our undersmoothed IPW estimators ought to be unbiased and efficient.

We consider undersmoothing criteria including the minimizer of D_{CAR} (Equation (5)) and the score-based method (Equation (6)). Throughout, we use the `hal9001` `R` package (Coyle et al., 2020; Hejazi et al., 2020), considering basis functions for up to all two-way interactions in estimating the propensity score and outcome models. For comparison, we construct propensity score estimates using a cross-validated HAL and a (parametric) logistic regression model with main effect terms for W_1 and W_2. All models were fit using 15-fold cross-validation. All numerical experiments were performed using the `R` language and environment for statistical computing (R Core Team, 2022).

Figures 1 and 2 display the results for scenarios 1 and 2, respectively. The IPW estimators using undersmoothing of the HAL outperform those based on cross-validated HAL in terms of both bias and efficiency, producing similar results as the IPW estimators based on correctly specified parametric models of the propensity score.

The first row of each figure presents the bias and the cross-validated mean of D_{CAR} (both scaled by $n^{1/2}$) of the corresponding estimators, where the latter is the objective function in Equation (5), and expected to be nearly zero for estimators that solve the EIF equation. While the scaled bias and the cross-validated mean of D_{CAR} of the cross-validation-based selector diverges (triangle), the undersmoothed HAL and the correctly specified parametric models perform similarly. In terms of coverage, D_{CAR}-based criterion achieves the nominal coverage rate of 95%, even for smaller samples sizes, whereas the cross-validation-based estimator (triangle) yields a poor coverage rate of ≈50%. The score-based undersmoothing selectors perform reasonably well, producing IPW estimators with coverage rates ≈90% for $n = 1000$ and ≈95% at larger sample sizes ($n \geq 5000$). In scenario 2, where the parametric model of the propensity score is misspecified, the parametric IPW estimator performs poorly, resulting in IPW estimators with coverage rates tending to zero asymptotically. In the same scenario, the score-based selector performs as well as the D_{CAR}-based selector producing estimators with coverage rates ≈95% for all the sample sizes considered. For both scenarios, we additionally report the selected tuning parameter λ based on both the global cross-validation and undersmoothing selectors. Our results illustrate that, as sample size increases, the selected value of the tuning parameter stabilizes. Importantly, this observation implies that the undersmoothing procedure does not lead to violations of the Donsker class assumption. Figures S6 and S7 in Section S4 of the Supporting Information show how the proposed estimator changes as a function of the tuning parameter λ. The U-shaped plots indicate that our proposed criteria perform well in terms of both scaled bias and the coverage of 95% Wald-style confidence intervals. Figures S8 and S9 in Section S4 of the Supporting Information show that when the cross-fitting is not used, our λ selector criteria tend to undersmooth too much resulting in inferior performance compared with the corresponding estimators with cross-fitting (Figures 1 and 2).

We provide additional simulation studies in Section S3 of the Supporting Information, in which we examine the relative performance of our proposed estimators under differing outcome and propensity score models. In Section S3, we also compare our proposed estimators with various augmented IPW estimators (Robins et al., 1994; Tsiatis, 2007; van der Laan and Robins, 2003). Our results suggest that, when at least one of the nuisance parameters is misspecified, bias of the doubly robust estimators tends to zero at a slower rate than $n^{-1/2}$, whereas the bias of the proposed IPW estimators tends to zero faster. Moreover, in relatively limited sample sizes, the D_{CAR}-based and the score-based IPW estimators can outperform the doubly robust estimators even when both nuisance parameters are consistently estimated. Notably, the score-based estimator does not even require any knowledge about the form of the EIF.

6 | EMPIRICAL ILLUSTRATION

6.1 | Overview and problem setup

We now apply our proposed estimation strategy to assessing the effect of smoking cessation on weight gain, using a subset of data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). As per Hernán and Robins (2020), the NHEFS was jointly initiated by the National Center for Health Statistics and the National Institute on Aging, in collaboration with several other agencies of the United States Public Health Service. The study was designed to investigate the impact of a variety of clinical, nutritional, and behavioral factors on health outcomes including morbidity and mortality. The subset of the NHEFS data we consider totaled $n = 1566$ cigarette smokers, all between the ages of 25 and 74; the data are publicly available. Each individual must have been present for a baseline visit and a follow-up visit roughly 10 years later. Individual weight gain was measured as a difference between baseline body weight and body weight at a follow-up visit; moreover, individuals were classified as having been in the treatment group if they reported having quit smoking prior to the follow-up visit and in the control group otherwise. Hernán and Robins (2020) caution that this subset of the NHEFS data could suffer from selection bias. As correcting for such a bias is tangent to the illustration of our analytic approach, we forego standard corrections, warning of this as a caveat of our demonstration. In practice, we advocate the use of our strategy in tandem with censoring or selection bias corrections, for example, imputation or inverse probability of censoring reweighting (e.g., Carpenter et al., 2006; Hernán & Robins, 2020; Seaman et al., 2012).

6.2 | Estimation strategy

We consider estimating the average treatment effect (ATE) of smoking cessation on weight gain in this subset of the NHEFS cohort ($n = 1566$). A fairly rich set of baseline covariates—including sex, race, age, highest degree of formal education, intensity of smoking, years of smoking, exercise habits, indicators of an active lifestyle, and weight at study onset—were considered as potential baseline confounders of the relationship between smoking cessation and weight gain. Constructing IPW estimators for the ATE requires estimation of the propensity score, to model the conditional probability of smoking cessation given potential baseline confounders. An IPW estimator for the ATE of smoking cessation may be constructed based on distinct estimators of the respective treatment-specific counterfactual means. We compare estimates of the ATE based on both parametric and nonparametric strategies for estimating the propensity score, including

(i) logistic regression with main terms for all baseline covariates;
(ii) logistic regression with main terms for all baseline covariates and with quadratic terms for age, smoking intensity, years of smoking, and baseline weight; and
(iii) the HAL with basis functions for all terms up to and including four-way interactions between the baseline covariates, fit with fivefold cross-validation.

The series of HAL propensity score estimators was constructed by weakening the restriction placed on the L_1-norm by a fit of the global cross-validation selector.

6.3 | Results

We apply each of the IPW estimators to recover the ATE of smoking cessation on weight gain, controlling for possible confounding by the baseline covariates previously enumerated. Table 1 summarizes the results. The unadjusted estimate is merely the difference of the mean observed outcomes between treated and untreated individuals. Generally, estimates of the ATE were similar across the two classes of IPW estimators. When the propensity score was estimated via a main terms logistic regression model, the estimate was 3.32 (CI: [2.15, 4.49]); likewise, when a logistic regression model with several quadratic terms was used, the estimate was 3.42 (CI: [2.24, 4.61]). By contrast, our cross-fitted (fivefold) nonparametric IPW estimators based on the HAL produced estimates of 3.20 (CI: [2.00, 4.41]) and 3.25 (CI: [2.03, 4.47]), for the cross-validation-based and D_{CAR}-based variants, respectively. As the form of the canonical gradient is readily known for the ATE, in this case, the D_{CAR}-based estimator provides the most reliable estimate. We note that the D_{CAR}-based estimate of the ATE is lower in magnitude than those recovered by parametric methods, suggesting that the impact of smoking cessation on weight gain may perhaps be lower than suggested by parametric propensity score estimation techniques, which

TABLE 1 NHEFS data: treatment effect estimates from different IPW estimators

Estimator	Lower 95% CL	Estimate	Upper 95% CL
HAL (undersmoothed, truncated)	2.24	3.48	4.72
HAL (undersmoothed, minimal truncation)	2.03	3.25	4.47
HAL (global cross-validation)	2.00	3.20	4.41
GLM (main terms only)	2.15	3.32	4.49
GLM (w/ quadratic terms)	2.24	3.42	4.61
Unadjusted (intercept model)	1.49	2.54	3.5

are more prone to model misspecification bias than our proposed approach.

Figure S10 shows how the treatment effect estimates change as a function of the L_1-norm. Table S2 presents the standardized differences as a measure of covariate balance (Greifer, 2021), and Figure S11 depicts the overlap between the HAL-based and parametric model-based propensity score estimates. It has been suggested that imbalance should be considered potentially important if the absolute standardized difference is greater than 0.2 (Austin, 2009). Although HAL results in absolute standardized differences less than 0.2, the values are slightly higher than the corresponding logistic regression fits. This is because parametric models exactly solve the score equations corresponding to the specified model, so these models attain a better balance in finite samples for the covariates included in the model. In contrast, HAL approximately (up to $O_p(n^{-2/3})$) solves score equations corresponding to relatively high-dimensional *derived* features. A major drawback of parametric modeling is the potential failure to balance complex functions of covariates (e.g., second-order terms) that are not specified in the model, resulting in misspecification (see Section S3.4 of the Supporting Information). Moreover, since our proposed estimators achieve efficiency by solving the EIF, the type of balance achieved may not be amenable to measurement by the marginal covariate balance techniques in popular use. In our example, the resultant ATE estimates are similar in scale (and all very different from the unadjusted estimate), further suggesting that important confounders have been properly adjusted for in all cases.

7 | DISCUSSION

We have proposed a class of nonparametric IPW estimators in which the weighting mechanism is estimated via undersmoothing of the HAL regression function. A particularly interesting application of the proposed approach is in settings in which closed-form representations of the EIF of the target parameter of interest are intractable, such as the bivariate survival probability in bivariate (or, in general, d-variate) right-censored data (van der Laan, 1996) or the mean or truncated mean survival time in interval-censored data (e.g., Chapter 8 of van der Laan & Rose, 2018). In these two example problems, the proposed method can be leveraged to achieve efficient estimation based on the readily available IPW estimators, by undersmoothing the HAL fits of the probability of censoring. The development of asymptotic linearity of the resultant IPW estimators and the corresponding undersmoothing criteria merits further investigation. Notably, there are several key differences between developing undersmoothed plug-in and nonplug-in estimators, including the techniques used to prove their theoretical properties and the required assumptions for undersmoothing of nuisance function estimators, which we discuss in much greater detail in Section S5 of the Supporting Information. One may obviate the need for truncation in finite samples by enforcing the positivity restriction $\min(G_n, 1 - G_n) > \delta$ in HAL fit. This is an interesting approach from both methodological and practical perspective that could motivate future research.

ACKNOWLEDGMENTS

We thank David Benkeser for helpful discussions and practical insights. This work was partially supported by the National Institute on Drug Abuse, the National Institute on Alcohol Abuse and Alcoholism, the National Institute of Neurological Disorders and Stroke, the National Institute of Allergy and Infectious Diseases (award no. R01-AI074345), and the National Science Foundation (award no. DMS-2102840).

DATA AVAILABILITY STATEMENT

The National Health and Nutrition Examination Survey (NHEFS) data that support the findings in this paper are openly available at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at http://re3data.org/.

ORCID

Ashkan Ertefaie https://orcid.org/0000-0003-2611-9512
Nima S. Hejazi https://orcid.org/0000-0002-7127-2789

REFERENCES

Austin, P.C. (2009) Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics – Simulation and Computation*, 38(6), 1228–1234.

Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

Benkeser, D., Carone, M., van der Laan, M.J. & Gilbert, P.B. (2017) Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4), 863–880.

Benkeser, D. & van der Laan, M.J. (2016) The highly adaptive lasso estimator. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 689–696.

Bibaut, A.F. & van der Laan, M.J. (2019) Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*.

Cai, W. & van der Laan, M. (2019) Nonparametric bootstrap inference for the targeted highly adaptive lasso estimator. *arXiv preprint arXiv:1905.10299*.

Cao, W., Tsiatis, A.A. & Davidian, M. (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3), 723–734.

Carpenter, J.R., Kenward, M.G. & Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 571–584.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. & Newey, W. (2017) Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261–65.

Coyle, J.R., Hejazi, N.S. & van der Laan, M.J. (2020) hal9001: the scalable highly adaptive lasso. R package version 0.2.7.

Geman, S. & Hwang, C.-R. (1982) Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10, 401–414.

Gill, R.D., van der Laan, M.J. & Wellner, J.A. (1995) Inefficient estimators of the bivariate survival function for three models. In: *Annales de l'IHP Probabilités et Statistiques*, volume 31, pp. 545–597.

Greifer, N. (2021) cobalt: covariate balance tables and plots. R package version 4.3.1.

Hejazi, N.S., Coyle, J.R. & van der Laan, M.J. (2020) hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*, 5, 2526.

Hernán, M.A. & Robins, J.M. (2020) *Causal inference: what if*. Boca Raton, FL: CRC.

Hirano, K., Imbens, G.W. & Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Kang, J.D. & Schafer, J.L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523–539.

Klaassen, C.A. (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15, 1548–1562.

Qiu, H., Luedtke, A. & Carone, M. (2020) Universal sieve-based strategies for efficient estimation using machine learning tools. *arXiv preprint arXiv:2003.01856*.

R Core Team (2022) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846–866.

Rotnitzky, A., Robins, J.M. & Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444), 1321–1339.

Seaman, S.R., White, I.R., Copas, A.J. & Li, L. (2012) Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1), 129–137.

Tsiatis, A. (2007) *Semiparametric theory and missing data*. New York, NY: Springer.

van der Laan, M.J. (1996) Efficient estimation in the bivariate censoring model and repairing NPMLE. *Annals of Statistics*, 24(2), 596–627.

van der Laan, M.J. (2014) Targeted estimation of nuisance parameters to obtain valid statistical inference. *International Journal of Biostatistics*, 10(1), 29–57.

van der Laan, M.J. (2017) A generally efficient targeted minimum loss-based estimator based on the highly adaptive lasso. *International Journal of Biostatistics*, 13(2)

van der Laan, M.J., Benkeser, D. & Cai, W. (2019) Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *arXiv preprint arXiv:1908.05607*.

van der Laan, M.J. & Bibaut, A.F. (2017) Uniform consistency of the highly adaptive lasso estimator of infinite-dimensional parameters. *arXiv preprint arXiv:1709.06256*.

van der Laan, M.J. & Bibaut, A.F. (2017) Uniform consistency of the highly adaptive lasso estimator of infinite dimensional parameters. *arXiv preprint arXiv:1709.06256*.

van der Laan, M.J. & Robins, J.M. (2003) *Unified methods for censored longitudinal data and causality*. Cham: Springer.

van der Laan, M.J. & Rose, S. (2018) *Targeted learning in data science*. Cham: Springer.

Vermeulen, K. & Vansteelandt, S. (2015) Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511), 1024–1036.

Vermeulen, K. & Vansteelandt, S. (2016) Data-adaptive bias-reduced doubly robust estimation. *International Journal of Biostatistics*, 12(1), 253–282.

Wasserman, L. (2006) *All of nonparametric statistics*. England: Springer Science & Business Media.

Zheng, W. & van der Laan, M.J. (2011) Cross-validated targeted minimum-loss-based estimation. In van der Laan, M.J. & Rose, S. (Eds.), *Targeted Learning* (pp. 459–474). Berlin: Springer.

SUPPORTING INFORMATION

Web appendices, tables, and figures referenced in Sections 1, 2.3, 3, 4.1, 5, 6.3, and 7 are available with this paper at the Biometrics website on Wiley Online

Library: (S1) Sectional variation norm; (S2) Proofs of Lemma 1 and Theorem 1; (S3) Additional Simulation Studies; (S4) Additional tables and figures; and (S5) Undersmoothing plug-in and nonplug-in estimators. An R code implementing our proposed approach is available with this paper and at https://github.com/nhejazi/pub_ipwhal_biometrics including code for both simulation studies and the real data analysis.

Figure S1: Relative performance of a simple unadjusted estimator and cross-fitted inverse-probability-weighted estimators based on the undersmoothing selectors and the cross-validation selector, in a randomized controlled trial

How to cite this article: Ertefaie, A., Hejazi, N.S., & van der Laan, M.J. (2023) Nonparametric inverse-probability-weighted estimators based on the highly adaptive lasso. *Biometrics*, 79, 1029–1041. https://doi.org/10.1111/biom.13719